Universidad Autónoma de Madrid
Escuela politécnica superior

Doble grado en Ingeniería informática y Matemáticas

Trabajo Fin de Grado

MODELOS PROBABILÍSTICOS GRÁFICOS PARA PREDICCIÓN DE PERFILES CRIMINALES

Autor: Mikel Sarriguren Ozcariz
Tutor: Daniel Ramos Castro

Junio 2018

MODELOS PROBABILÍSTICOS GRÁFICOS PARA PREDICCIÓN DE PERFILES CRIMINALES

Autor: Mikel Sarriguren Ozcariz
Tutor: Daniel Ramos Castro

Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio 2018

Resumen

Resumen

La criminología es una disciplina que tiene por objeto el estudio del criminal con relación al crimen con el objetivo de entender las distintas motivaciones que lo llevaron a cometer sus actos. Una de sus ramas es el perfilado criminal, cuyo objetivo es definir los llamados *perfiles criminales* estableciendo los patrones de conducta o características que comparten ciertos criminales a partir de, normalmente, un conjunto reducido de datos del autor, la víctima, o el escenario de un crimen.

En este contexto, y para poder llevar a cabo dichas tareas, se han utilizado tradicionalmente modelos y técnicas estadísticas tales como algoritmos de agrupamiento, regresión, o agrupación categórica o cuantitativa. Durante los últimos años se han comenzado a utilizar algunos modelos más complejos, los llamados modelos probabilísticos gráficos, como las redes bayesianas, que proporcionan una mayor capacidad holística y funcional que los modelos anteriormente mencionados que se venían usando.

Este tipo de nuevos modelos, y más en concreto las redes bayesianas, son especialmente interesantes aplicadas a estas disciplinas, ya que permiten no solo asociar todas las variables de un mismo caso en un solo modelo sino además predecir variables no observadas o desconocidas en un caso a partir de las variables que sí son conocidas o se han podido observar, como pueden ser las variables relacionadas con el autor de un crimen sin resolver a partir de los datos observados del escenario o de la víctima.

En el presente trabajo de fin de grado nos centraremos en estudiar los diferentes algoritmos de entrenamiento e inferencia con redes bayesianas para el contexto explicado. A partir de una base de datos que recoge variables de casos de agresiones sexuales, trataremos de predecir las variables del autor del crimen.

Palabras Clave

Modelo gráfico, red bayesiana, inferencia, predicción de variables.

Índice general

Índice de Figuras

Índice de Tablas

1

Introducción

1.1. Motivación del proyecto

Durante los últimos años en el marco de la criminología, y en relación con las Fuerzas y Cuerpos de Seguridad del Estado (en adelante FCSE) se vienen realizando numerosos esfuerzos en la generación de perfiles criminales relacionados con agresiones sexuales. Durante el estudio de estos casos los expertos de las FCSE recogen gran cantidad de datos sobre el *modus operandi* del autor, incluyendo entre estos toda la información relevante sobre la forma de actuar del agresor, el entorno donde se produjo el delito o información acerca de la víctima. No obstante, no son estas, sino las llamadas *variables de autor* relacionadas con el agresor (tales como edad, procedencia o características), normalmente desconocidas, las que interesan a los investigadores para facilitar la resolución del caso.

A partir de toda esta información recopilada por las FCSE ha estado trabajando el Instituto de las Ciencias Forenses y de la Seguridad de la Universidad Autónoma de Madrid (en adelante ICFS-UAM). Gracias a todos estos datos, y a un trabajo previo de categorización y selección de variables llevado por expertos en el área de las ciencias forenses, psicología y criminología, el ICFS-UAM ha sido capaz de generar una base de datos que refleja de forma realista la información acerca de estos casos. Esta base de datos nos permitirá trabajar sobre los casos recopilados con modelos probabilísticos gráficos para buscar relaciones entre los dos grupos de variables. El objetivo principal será intentar predecir correctamente datos sobre las *variables de autor* a partir de los datos observados en las variables de *modus operandi*.

Para llevar a cabo esta tarea es necesario buscar un modelo estadístico capaz de realizar dos funciones principales. Por un lado es necesario un modelo que sea capaz de establecer relaciones holísticas entre todas las variables de nuestra base de datos, de esta manera podremos obtener información sobre posibles variables desconocidas a partir de otras dadas, los modelos gráficos resultan ser una buena solución para ello. Por otro lado, debe tratarse de un modelo que permita realizar predicciones de alguna de sus variables a partir de observaciones totales o parciales de las demás, en nuestro caso trataremos de predecir las variables desconocidas *variables de autor* a partir de las variables de *modus operandi*. Por estas y por otras razones que se comentarán en el apartado correspondiente fueron elegidas las Redes Bayesianas (en adelante RB) para modelar nuestro problema.

1.2. Objetivos y enfoque

Los objetivos del presente Proyecto son los siguientes:

1. Hacer un estudio teórico de los modelos gráficos probabilísticos, más en concreto de las RB y su aplicación en problemas parecidos a los que nos enfrentaremos en el trabajo.
2. Generación y entrenamiento de Redes Bayesianas en R utilizando la base de datos generada por el IFCS-UAM.
3. Estudio empírico de las capacidades predictivas de las RB implementadas utilizando los datos del IFCS-UAM.
4. Estudio de las RB generadas y su funcionalidad utilizando la herramienta HuginLite.

Tanto la implementación de las redes como las pruebas y experimentos serán realizados en el lenguaje de programación R gracias a su librería catNet. El lenguaje R es uno de los más utilizados en el campo del análisis de datos y cálculo científico debido a su simplicidad.

1.3. Organización del documento

El documento se ha organizado del siguiente modo

- Capítulo 2 : Modelos Gráficos y Redes Bayesianas.
  - Breve introducción a las estructuras probabilísticas que usaremos en los experimentos de este trabajo incluyendo un pequeño ejemplo.
- Capítulo 3: Sistema, diseño y desarrollo.
  - En este apartado se definirán y explicarán las herramientas utilizadas tanto para la implementación de las Redes Bayesianas como de las pruebas realizadas. También se detallarán los algoritmos utilizados para el entrenamiento y predicción de las redes.
- Capítulo 4: Experimentos realizados y resultados.
  - En este capítulo pasaremos de la teoría a la práctica y detallaremos todos los experimentos realizados con el objeto de comprobar empíricamente la utilidad de las metodologías y modelos estudiados en los capítulos anteriores.
- Capítulo 5: Conclusiones y trabajo futuro.
  - Por último, detallaremos las conclusiones que hemos podido sacar después de realizar los experimentos y propondremos posibles tareas para realizar en el futuro siguiendo el hilo del proyecto.

2

Modelos gráficos y redes bayesianas

2.1. Introducción a los modelos gráficos

Los modelos gráficos probabilísticos (en adelante MG) [*Koehler y Friedman, 2009*][1] son estructuras que combinan incertidumbre (probabilidades) y estructura lógica para representar, de forma robusta, complejos fenómenos del mundo que nos rodea. Los MG han experimentado un auge de interés a lo largo de las últimas dos décadas debido tanto a su flexibilidad y potencial de representación como a la mejora en la capacidad de aprender e inferir sobre redes de gran tamaño.

Debido a esto los MG se han convertido en una herramienta extremadamente popular a la hora de crear modelos de incertidumbre (probabilidades). Apoyados en la teoría de la probabilidad y la teoría de grafos nos proporcionan una gran forma de lidiar tanto con la incertidumbre como con la complejidad de algunos modelos. Los dos tipos de MG más populares y utilizados son las cadenas de Markov y las redes bayesianas (en adelante RB). En los siguientes puntos centraremos nuestros esfuerzos en explicar el funcionamiento y la utilidad de estas últimas, ya que son las redes que se han utilizado para realizar todo lo relativo a este trabajo.

2.2. Redes bayesianas

Las RB son una herramienta matemática que nos proporciona una relación visual para una función de densidad de probabilidad conjunta a un grupo de variables de un mismo problema que presente incertidumbre. Esta función de densidad conjunta es todo lo necesario para modelar cualquier problema de forma completa, ya que por un lado podemos obtener las probabilidades individuales de cada variable a partir de la probabilidad conjunta utilizando operaciones de marginalización, y por otro lado podemos obtener las probabilidades condicionadas a partir de la conjunta y las individuales mediante la aplicación directa del teorema de Bayes .

Las RB se representan mediante un grafo dirigido acíclico. Cada uno de los nodos representa una de las variables aleatorias que queremos estudiar, mientras que las conexiones (o flechas), representan una influencia de una sobre la otra, dependiendo de la dirección que tome dicha conexión. Esta representación nos permite, incluso sin tener conocimientos matemáticos o probabilísticos avanzados, hacernos una idea intuitiva de las variables que pueden proporcionar

algún tipo de información sobre otras. Esto es, si dos variables están conectadas, el conocimiento del valor de una de ellas nos proporcionará información de las probabilidades que hay de que la otra tome un valor determinado, ya sea mediante inferencia evidencial (sentido opuesto a la dirección de la conexión) o causal (en el sentido de la conexión, en este caso es aplicación directa de las probabilidades condicionadas).

A continuación se expone una lista de algunas de las propiedades de interés por las que se han elegido las RB para este proyecto:

- Se trata de una estructura gráfica bastante explícita de fácil comprensión, lo cual supone una gran ventaja tanto a la hora de visualizar el trabajo llevado a cabo como a la hora de exponerlo y explicarlo.
- Es una forma sencilla de modelar los problemas a los que nos enfrentamos en este trabajo. Simplemente basta con asignar una variable a cada nodo, sin necesidad de tener que agruparlas o emparejarlas de algún modo concreto.
- Tiene capacidad de aprendizaje. Esto es, a partir de un conjunto de datos de entrada, la red es capaz de entrenarse (concepto detallado en los siguientes capítulos), ajustando y estableciendo las relaciones entre las variables en función de los datos observados.
- Una vez entrenada podemos usarla para hacer predicciones sobre variables con datos desconocidos a partir de otras conocidas.

Sin embargo, también tienen algún aspecto negativo que limitará algunos temas a la hora de trabajar con ellas:

- Complejidad algorítmica. Si bien el concepto de RB es relativamente simple y fácil de visualizar, las RB son estructuras computacionalmente complejas, y los algoritmos de entrenamiento y predicción son bastante sofisticados. Esto supone una limitación a la hora de trabajar con un número elevado de variables u observaciones.
- Posibilidad de confusión con la interpretación. A diferencia de otros tipos de modelos comunes en los que también se usan grafos dirigidos, en las RB una conexión entre dos variables no establece una relación causa/efecto, sino una asociación estadística.

Para entender mejor las RB, a continuación se presenta un ejemplo sencillo.

2.3. Ejemplo de red bayesiana sencilla

2.3.1. Representación

El siguiente ejemplo, obtenido de [*Murphy, 2001*][2] refleja de una forma sencilla el funcionamiento de una RB y las posibilidades que ésta nos ofrece. El ejemplo con la representación gráfica y sus tablas de probabilidad puede verse en la figura 2.1

El problema presentado estudia la probabilidad de que la hierba de un determinado lugar esté mojada o no en relación con la situación meteorológica (lluvia y nubosidad) y la actividad de un aspersor. Las variables que tendremos entonces serán las siguientes:

- *Cloudy (C)*: variable booleana, T si hay nubes en el cielo y F si no hay.
- *Rain (R)*: variable booleana, T si está lloviendo y F si no está lloviendo.

- *Sprinkler (S)*: variable booleana, T si el aspersor está en funcionamiento y F si está apagado.
- *WetGrass (W)*: variable booleana, T si la hierba está mojada y F si está seca.

Estas variables, como se puede suponer están estrechamente relacionadas entre ellas. Es normal suponer que si está lloviendo, la probabilidad de que la hierba esté mojada sea bastante alta, y lo mismo ocurrirá con el funcionamiento del aspersor.

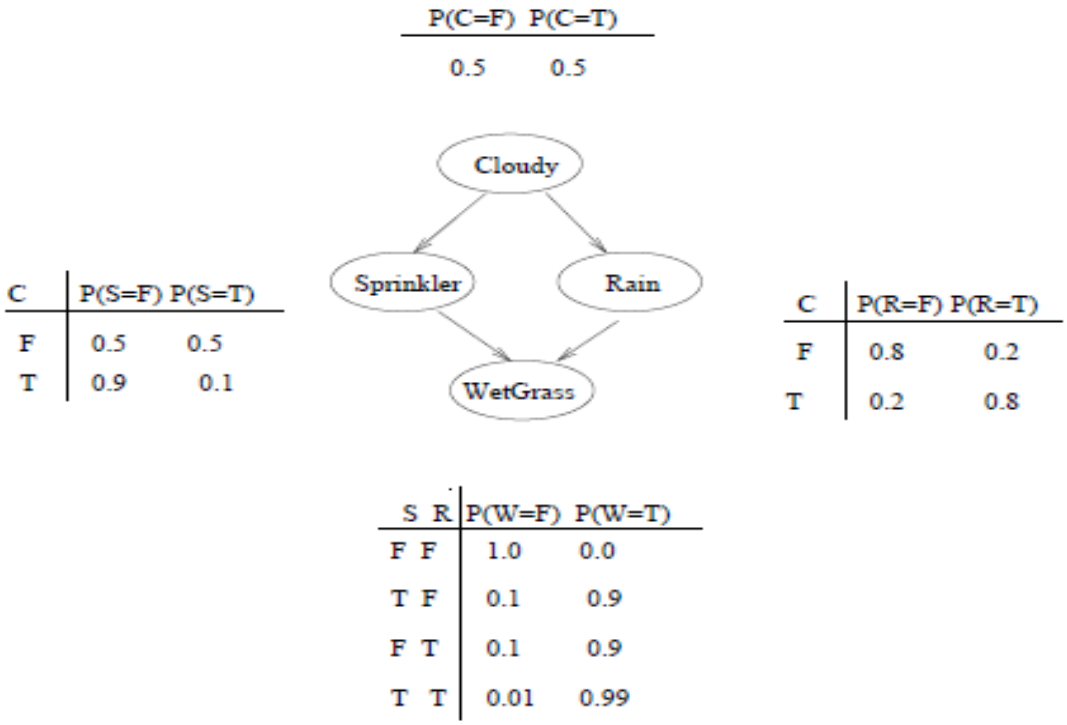

Figura 2.1: Ejemplo de red bayesiana (fuente: Murphy, 2001).

A continuación se explican los términos o conceptos más importantes referentes a las RB que serán utilizados con frecuencia en el documento.

- *Nodo padre/ancestro*: Un nodo A es padre de un nodo B si existe una conexión dirigida de A a B. A su vez, un nodo es ancestro si es padre de un nodo padre.
- *Tabla de probabilidades*: Todas las relaciones holísticas entre las probabilidades de las variables están reflejadas en las tablas de probabilidades. A cada nodo le corresponde una tabla y ésta determina las probabilidades de que el nodo tome un valor determinado conociendo los valores de sus nodos padres.
- *Probabilidad condicionada*: La probabilidad de un nodo A condicionado a B es la probabilidad de que A tome cierto valor conociendo el valor de B (siendo B padre de A).
- *Independencia condicional*: Dos nodos A y B son condicionalmente independientes dado un tercer nodo C, si la probabilidad de que A y B tomen unos valores concretos, se da de forma independiente dado el valor de C.
- *Propiedad local de Markov*: Es la propiedad fundamental que define la actividad de toda RB. Esta propiedad establece que para cierta variable o nodo, si conocemos o podemos

observar el valor de sus nodos padres entonces su valor es independiente de todos sus ancestros.

2.3.2. Inferencia

Llamamos inferencia a la acción de deducir o estimar las probabilidades de los valores que no conocemos a partir de valores de variables que si conocemos o son observables. En nuestro ejemplo un caso de inferencia sería establecer la probabilidad de que el aspersor esté encendido sabiendo que la hierba está mojada, o a la inversa, determinar la probabilidad de que la hierba esté mojada sabiendo que el aspersor está encendido.

Atendiendo a esto podemos distinguir dos tipos de inferencia:

- Inferencia evidencial, diagnosis o *bottom-up reasoning*. En este caso intentamos estimar las probabilidades de abajo a arriba, o lo que es lo mismo, a partir de una evidencia (la hierba está mojada) intentaremos establecer la probabilidad de cual ha podido ser la causa de ello (por ejemplo que el aspersor esté encendido).
- Inferencia causal, predicción o *top-down reasoning*. En este caso observando el valor de una variable (el aspersor está en funcionamiento) intentaremos estimar la probabilidad que tiene de que cause cierto valor en sus nodos hijos (por ejemplo que la hierba esté mojada).

2.3.3. Aprendizaje

Denominamos aprendizaje al proceso de establecer, a partir de un conjunto de datos de entrenamiento, los parámetros de una red bayesiana que se corresponda con los datos introducidos. En las RB que usaremos distinguiremos dos tipos distintos de aprendizaje.

- *Aprendizaje estructural*. Este aprendizaje tiene por objeto el establecer la estructura del grafo y las conexiones entre los nodos. Para llevarlo a cabo se suelen utilizar algoritmos de un alto coste computacional que requieren estrategias de computación paralela. Para lidiar con este problema, durante las pruebas realizadas en este proyecto se ha establecido un orden determinado para los nodos en el entrenamiento, causando que los nodos de orden más bajo tiendan a ser hijos de los nodos de mayor orden. Con esto conseguimos que el proceso de aprendizaje sea mucho menos complejo computacionalmente a costa de imponer un orden establecido en la red, lo que supone restringir en cierta manera los posibles grafos que se pueden generar.
- *Aprendizaje de parámetros*. El aprendizaje de parámetros parte de una red ya estructurada bien por entrenamiento o bien establecia por expertos, y da valor a todos los parámetros necesarios para calcular las tablas de las probabilidades de cada nodo. En casos como al que nos enfrentamos en este trabajo, donde las variables son discretas, la distribución probabilística idónea que usaremos es la multinomial, que simplemente asigna a cada valor una probabilidad proporcional a la frecuencia de aparición de dicho valor.

3

Sistema, diseño y desarrollo

3.1. El lenguaje R, catNet y HuginLite

A continuación describiremos los lenguajes y herramientas que se han utilizado para la realización de los experimentos llevados a cabo.

Para este proyecto se necesitaba un lenguaje con ciertos requisitos que son enumerados a continuación:

- Debía de tratarse de un lenguaje con un claro enfoque matemático, más concretamente probabilístico, ya que la gran mayoría de nuestras operaciones iban a ser de este carácter.
- Facilidad a la hora de crear gráficas y/o tablas. Ya que gran parte del trabajo consta de analizar los resultados obtenidos en la predicción, era necesario un lenguaje que fuera capaz de analizar y plasmar estos datos en gráficas de forma sencilla y fluida, sin necesidad de usar programas externos.
- Por último, el requisito más importante era contar con alguna librería que permitiera tanto implementar y entrenar de forma sencilla nuestras RB a partir de la base de datos, como predecir valores no definidos a partir de los demás. La librería de *R* catNet era la idónea para ello.

Por todo ello se determinó que el lenguaje idóneo debía ser *R*.

Para el análisis y estudio de las redes generadas mediante el código en *R* se ha utilizado el software *HuginLite*, para lo cual ha sido necesario implementar en *R* una función que adapte los datos relativos a la red que queremos estudiar en *R* a los ficheros para importación de HuginLite con su correspondiente formato.

3.1.1. El lenguaje de programación R

R es un entorno y lenguaje de programación con un claro enfoque al análisis estadístico. En la actualidad se trata de uno de los lenguajes más utilizados en investigación por la comunidad estadística y matemática. Uno de sus puntos fuertes es la facilidad para crear y compartir librerías

o paquetes con distintas funcionalidades, lo que posibilita que exista una gran diversidad de funcionalidades ya implementadas que pueden ser incluidas para el presente proyecto.

3.1.2. El paquete catNet para R

A continuación se expone un pequeño resumen de lo que el paquete catNet nos ofrece y su importancia a la hora de realizar las pruebas de este trabajo. Los datos han sido obtenidos de [*Balov y Saltzman, 2017*][3].

La librería catNet dispone de una estructura de datos en *R* para almacenar toda la información relativa a una RB. Además, proporciona una serie de funciones de predicción y aprendizaje. Estas últimas basadas en la búsqueda exhaustiva mediante el aumento de la complejidad de las redes que ajusta los parámetros utilizando el criterio de máxima verosimilitud *MLE (Maximum-Likelihood Estimation)* y el algoritmo esperanza-maximización *EM (Expectation-Maximization)*.

No obstante la librería tiene el inconveniente de que no permite trabajar con variables aleatorias de tipo continuo, sino solo con variables discretas categóricas. Por suerte, esto no ha resultado ser un inconveniente ya que todas las variables con las que trabajamos en este proyecto son de tipo categórico con variables distribuidas de forma multinomial.

Por desgracia catNet no dispone de ninguna funcionalidad para visualizar las redes obtenidas, y por ello hay recurrir a herramientas externas para poder hacerlo. En el caso del presente trabajo se han utilizado *graphviz* y *HuginLite*, siendo este segundo el que ha permitido estudiar al máximo las RB que resultaban del aprendizaje, ya que no solo aporta una representación gráfica sino que permite cargar las *tablas de probabilidades* y analizarlas.

3.1.3. El programa HuginLite

Como hemos visto, catNet no dispone de ninguna herramienta para visualizar o analizar las redes que se obtienen del entrenamiento. Es por esto por lo que se ha recurrido a la herramienta *HuginLite*. Este software permite cargar las RB ofreciendo una representación gráfica. Adicionalmente proporciona una interfaz de ésta que permite estudiar y analizar las *tablas de probabilidades* de forma interactiva, dejando establecer o fijar los valores de los nodos y ver las probabilidades condicionadas a ese valor. Esto nos da una visión general bastante satisfactoria del funcionamiento de la red y da la posibilidad de analizar en cada caso la causa de los resultados obtenidos (si es algo no esperado).

Por desgracia los datos para *HuginLite* no pueden ser cargados directamente desde *R*, sino que hay que importarlos desde ficheros con un formato específico. Para pasar de nuestros datos en *R* a los ficheros que requiere *HuginLite* se ha creado el paquete *catNetToHugin*.

3.2. Esquemas de agrupamiento de variables

Se llaman esquemas de agrupamiento a las distintas formas que existen de agrupar las variables disponibles para el entrenamiento en problemas de predicción. Esto es de gran importancia ya que, como es lógico, las variables se comportan o asocian de forma distinta en función de las variables disponibles, dando lugar a resultados distintos como comprobaremos en los experimentos. Para este trabajo se han utilizado dos esquemas distintos de agrupamiento que son los siguientes:

- *Agrupamiento separado*. En un primer lugar se ha pensado que buscar la red mejor o más especializada a la hora de predecir cada una de las *variables de autor* sería interesante para

ver el comportamiento por separado de éstas y analizar las variables de *modus operandi* más relacionadas con cada una de ellas. En este esquema se ha entrenado una RB para cada una de las *variables de autor*. Para ello, cada RB entrenada parte de una de las *variables de autor* a predecir y se añaden otras variables de *modus operandi* según criterios que más tarde serán explicados. Así pues, las *variables de autor* se predicen cada una con un conjunto específico de variables, optimizando de esta forma las capacidades predictivas de la red para esta variable en concreto. Este tipo de esquemas son, como veremos más adelante, los que han conseguido mejores resultados a la hora de la predicción.

- *Agrupamiento simultáneo*. En este esquema se parte del conjunto de todas las *variables de autor* y se añaden otras variables de *modus operandi*. De este modo la red quedará entrenada con todas las variables implicadas en la predicción presentes a la vez. Una vez entrenada la RB ofrece la posibilidad de predecir las *variables de autor* como se prefiera, bien las tres que existen a la vez, de dos en dos o de una en una. Si bien esta sería la forma más adecuada de solucionar los problemas a los que nos enfrentamos, los resultados obtenidos han sido en general peores que los de agrupamiento separado, obteniendo resultados positivos en tan solo uno de los experimentos realizados con este esquema.

3.3. Algoritmos de entrenamiento y predicción de redes bayesianas en R

Distinguiremos entre dos esquemas de entrenamiento estructural diferentes:

- *Entrenamiento con orden pre-establecido*. En este caso, el algoritmo de entrenamiento cuenta con un orden dado para los nodos, siendo los nodos de órdenes más bajos los que van a tender a ser hijos. Esto quiere decir que tan solo valorará redes que cumplan esa jerarquía descartando todas las demás. Es por lo tanto de vital importancia hacer un estudio previo intensivo sobre el orden que queramos dar a dichas redes. Estos esquemas tienen la ventaja de ser relativamente sencillos en comparación con los que se explican a continuación.

- *Entrenamiento sin orden pre-establecido*. En este caso, el algoritmo de entrenamiento valora todas las posibles redes que se pueden llegar a formar con los nodos proporcionados y busca entre todas ellas la que mejor rendimiento otorga. Como es de suponer, este tipo de esquemas son muy complejos en comparación con los de orden pre-establecido, ya que se prueban infinidad de combinaciones distintas que ni se contemplan en el caso de disponer de un orden de nodos dado. Estos esquemas, en general, nos proporcionarán redes más sofisticadas que las de orden pre-establecido a costa de aumentar en gran medida la complejidad del entrenamiento.

Como ya habíamos visto anteriormente, el entrenamiento estructural de una red es un proceso computacionalmente muy costoso. Para reducir esta complejidad añadida a nuestras pruebas utilizaremos esquemas de entrenamiento de orden pre-establecido con el orden previamente estudiado y analizado. Esto nos permitirá obtener los datos de forma más eficiente a costa de imponer un orden en las redes exploradas por los algoritmos de entrenamiento.

A continuación se explica cómo se ha llevado a cabo el entrenamiento de parámetros de las RB en las pruebas realizadas.

Uno de los mayores problemas al que nos enfrentamos a la hora de llevar a cabo el entrenamiento de parámentros de cualquier modelo estadístico de predicción es la selección de los bloques de datos que serán utilizados en el entrenamiento y las pruebas.

Por un lado, estos bloques deben ser claramente distintos para evitar lo denominado sobreajuste u *overfitting*. Si esto ocurre es posible que la red entrenada funcione correctamente para predecir los datos que conoce (con los cuales se ha entrenado y se ha probado), pero en cambio tendrá grandes problemas a la hora de reconocer e intentar predecir casos que no ha visto antes. Es por tanto necesario buscar una división de los datos disponibles en estos dos bloques que garantice una estabilidad a la hora de predecir en la red entrenada.

Por otro lado, esta división se hace más complicada en problemas en los que el número de casos de los que se dispone es limitado o bajo, como es el nuestro. A menor número de casos en una base de datos más complicado será que estos representen de forma correcta todas las posibles situaciones del problema, con lo que se complica el entrenamiento y la red resultante no será capaz de predecir correctamente casos que no estén ahí reflejados.

Esta escasez de datos supone un problema adicional a la hora de seleccionar los bloques de entrenamiento y pruebas, ya que alguno de estos puede resultar insuficientemente grande para garantizar un buen aprendizaje de la red. Esto supone un inconveniente tanto para el bloque de pruebas, que necesita suficientes casos para garantizar robustez de los resultados de predicción obtenidos, como para el bloque de entrenamiento, que a menor número de casos disponible mayor será el número de situaciones desconocidas.

Para buscar solución a este problema usaremos una estrategia iterativa de entrenamiento conocida como *validación cruzada* [*R. Duda et al., 2000)*][4]. Mediante esta estrategia se realiza una división iterativa de datos que garantice abarcar todos ellos tanto para entrenamiento como para prueba. Para ello, en cada paso de la iteración, se dividen los datos disponibles en un conjunto de entrenamiento amplio y un conjunto de prueba con escasos datos (de ahora en adelante a este conjunto de prueba lo denominaremos *fold*). De tal modo que se predicen los casos del *fold* habiendo entrenado con el resto. En el siguiente paso de la iteración se selecciona un nuevo *fold*, de tal forma que cuando todos los casos de la base de datos han sido alguna vez parte del *fold* (han formado parte del grupo de pruebas) el proceso termina. En este punto se juntan los resultados obtenidos en cada iteración para obtener una medida de rendimiento global de todos los datos disponibles.

De esta forma garantizamos que los conjuntos de prueba y entrenamiento sean distintos mientras que estos son lo más grande posible para los datos disponibles.

Como es de suponer el tamaño del *fold* es de relevancia a la hora de obtener unos resultados mejores o peores. En el caso de *fold* de tamaño pequeño estaremos utilizando para el entrenamiento casi la totalidad de los datos, por lo que la probabilidad de encontrarse con una situación no conocida a la hora de predecir es menor. En este caso el rendimiento predictivo de la red será *optimista*, ya que en la práctica esto no tiene por qué ocurrir y los resultados obtenidos serán mejores. En el caso opuesto, por un razonamiento similar al anterior, tendremos un conjunto de entrenamiento reducido, por lo que la red no será capaz de reconocer todos los casos de prueba que se le presentan (que además son más numerosos), por eso decimos que el rendimiento de la red será *pesimista*.

Para este proyecto se valoró la posibilidad de utilizar la validación cruzada de dos formas distintas. En primer lugar una predicción optimista con tamaño de *fold* 1, es decir, el llamado *dejar uno fuera* (*Leave One Out*, *LOO*), y en segundo lugar una predicción más realista usando *fold* de tamaño 20. Por temas de complejidad algorítmica y tiempos de ejecución sólo se han podido llevar a cabo con garantías estos últimos.

4

Experimentos realizados y resultados

4.1. Base de datos y protocolo

En los siguientes puntos se describe la Base de Datos (en adelante BD) obtenida por el ICFS-UAM.

Durante meses el ICFS-UAM ha estado en contacto con las FCSE recogiendo información sobre delitos de agresiones sexuales que estos recopilaban. Estos datos fueron pre-procesados y volcados en una base de datos inicial que constaba de 471 casos, cada uno de ellos con información sobre 64 variables de *modus operandi* (MO) y 3 de características relacionadas con el autor (*variables de autor*), siendo estas últimas las de mayor importancia en la investigación pues son las que se tratan de obtener para resolver el caso. Las *variables de autor* contempladas son las siguentes:

- *Pais_ categ*: país de origen del autor del delito.
- *N_ antpolicial*: número o cantidad de antecedentes policiales del agresor.
- *Edad_ cat*: rango de edad del autor.

Como veremos próximamente, todas ellas (tanto de *modus operandi* como de autor) han sido previamente sometidas a un proceso de discretización que facilitará el uso de modelos gráficos y la predicción de variables, que como sabemos es el objetivo principal de nuestros experimentos.

4.2. Procesado de la base de datos y selección de variables

A continuación se detallan los pasos de procesado por los que han pasado los datos obtenidos hasta llegar a la base de datos definitiva con la que realizaremos los experimentos de este proyecto.

En primer lugar el personal experto del ICFS-UAM se ha encargado de discretizar todas las variables de las que se tenían datos. De esta forma se consigue simplificar en gran manera el problema, lo que nos permite aplicar modelos gráficos como las RB y llevar a cabo las predicciones previstas que serían imposibles de otro modo (recordemos que la librería utilizada catNet tan

solo permite trabajar con variables discretas). Este proceso de discretización ha sido llevado a cabo de tal manera que se ha intentado que para todas las variables presenten datos en todos sus valores, de esta manera se evitan problemas de robustez en las redes entrenadas. Como veremos más adelante este objetivo no se ha cumplido del todo ya que algunas de las variables tienen pocos datos para algunos de sus valores. Como veremos más adelante, en uno de los experimentos trataremos de seguir con esta discretización de las variables agrupando valores en algunas de las variables que tienen una distribución más asimétrica.

En segundo lugar, con el fin de hacer la base de datos con la que trabajaremos más manejable computacionalmente hablando, se ha hecho una selección de variables interesantes y que puedan aportar información relevante llevada a cabo por los expertos del ICFS-UAM.

En la tabla 4.1 se puede observar la estructura de la base de datos resultante tras este procesado, con la que llevaremos a cabo los experimentos. Como puede verse el número de variables se ha reducido a más de la mitad.

En los siguientes puntos se explicarán detalladamente las estrategias de selección de variables que se han utilizado para realizar los experimentos de este trabajo.

NOMBRE	AUTOR/MODUS O	TIPO	VALORES
Ag_serie	Modus Operandi	Categórica	[0,1]
Ant_noviolent	Modus Operandi	Booleana	[0,1]
Ant_sexu	Modus Operandi	Booleana	[0,1]
Ant_violent	Modus Operandi	Booleana	[0,1]
Arma	Modus Operandi	Categórica	[0,1,2]
Bajo_efecto	Modus Operandi	Booleana	[0,1]
Comp_rel	Modus Operandi	Categórica	[1,...,8]
Comp_sex	Modus Operandi	Categórica	[0,...,5]
Control	Modus Operandi	Categórica	[1,...,3]
CP	Modus Operandi	Categórica	[1,2]
Delito_cat	Modus Operandi	Categórica	[1,...,4]
Dia	Modus Operandi	Categórica	[1,...,7]
Distancia_ag_ca	Modus Operandi	Ordinal	[1,...,5]
Distancia_vict_c	Modus Operandi	Ordinal	[1,...,5]
Edad_cat	Autor	Ordinal	[1,...,5]
Edad_vict_cat	Modus Operandi	Ordinal	[1,...,5]
Esce_cat	Modus Operandi	Categórica	[1,...,5]
Espacio_ag	Modus Operandi	Categórica	[1,2]
Final	Modus Operandi	Categórica	[1,...,4]
Grado_ejec	Modus Operandi	Categórica	[1,2]
Grupo	Modus Operandi	Categórica	[0,1]
Lesiones	Modus Operandi	Categórica	[0,...,3]
Met_aprox	Modus Operandi	Categórica	[0,...,4]
Momento	Modus Operandi	Categórica	[1,2,3]
N_antpolicial	Autor	Ordinal	[0,...,2]
Objetos	Modus Operandi	Categórica	[0,...,7]
Pais_cat_vict	Modus Operandi	Categórica	[1,8]
Pais_categ	Autor	Categórica	[1,8]
Rel_area	Modus Operandi	Categórica	[1,...,4]
Sola	Modus Operandi	Booleana	[0,1]
Tipo_dia	Modus Operandi	Categórica	[1,2]
Uso_vehiculo	Modus Operandi	Booleana	[0,1]
Vict_minusv	Modus Operandi	Booleana	[0,1]
Vict_prost	Modus Operandi	Booleana	[0,1]

Cuadro 4.1: Variables seleccionadas y discretizadas por el ICFS. Se observan las variables que pertenecen al Modus Operandi (MO) y las variables de autor que se han intentado predecir con la RB.

4.2.1. Selección por patrones perdidos

En primer lugar se han tenido en cuenta las variables de la base de datos con un alto porcentaje de *valores perdidos* (valores no conocidos), ya que como puede suponerse esta variable pierde un gran valor informativo. Para ello se han eliminado las variables que tenían un porcentaje de valores perdidos superior a un umbral dado. El valor de este umbral se ha establecido en un 15 % para todas las pruebas.

4.2.2. Selección por distribución

En segundo lugar tendremos en cuenta la capacidad informativa de cada una de las variables por separado, para eso estudiaremos su distribución.

Por norma general, una variable proporcionará información más relevante si presenta una distribución suficiente en todos sus valores. Es decir, variables que tienen una distribución muy concentrada en unos pocos de sus valores o variables que disponen de muy pocos datos para algunos de sus valores tenderán a no ser muy informativas.

Para estudiar esta capacidad informativa de las variables se calculará su entropía, una medida típica sobre la dispersión o concentración de una distribución probabilística que nos da una idea de la cantidad de información que puede aportar dicha variable. Esta magnitud se define matemáticamente de la siguiente forma:

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i)$$

Donde X se refiere a una variable aleatoria, x_i a cada uno de los valores que puede tomar dicha variable y $p(x_i)$ la probabilidad de que se dé dicho valor.

Para su cálculo en *R* se dispone de la función *entropy* que la calcula proporcionando los datos necesarios. Con el fin de obtener su valor entre 0 y 1 sin depender del número de valores que puede tomar la variable aleatoria X se han normalizado los datos obtenidos de la siguiente manera:

$$H_N(X) = \frac{H(X)}{\log_2 N}$$

Donde N es el número total de registros de la base de datos y H_N es simplemente el nuevo nombre que se le otorga, significando que es la entropía normalizada para N casos en la base de datos.

Este tipo de selección de variables se ha utilizado en los entrenamientos con agrupación simultánea, ya que de esta forma iremos introduciendo en la RB las variables con más capacidad informativa. Posteriormente veremos que salvo en uno de los experimentos los resultados no han sido positivos usando este tipo de selección. En la gráfica 4.1 pueden observarse los valores de entropía ordenados de mayor a menor de las variables de *modus operandi*.

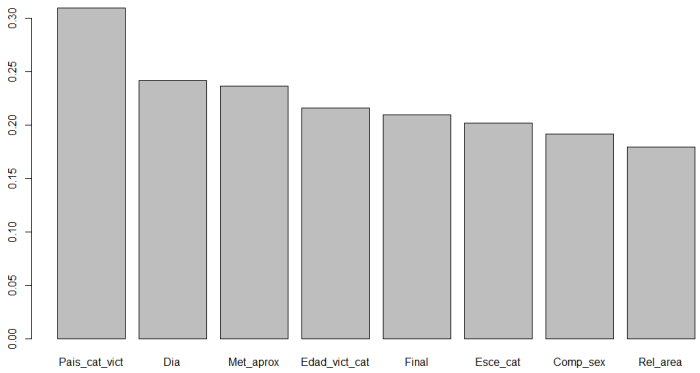

Figura 4.1: Valores de entropías para las variables de *modus operandi*

4.2.3. Selección por asociación

En tercer y último lugar tendremos en cuenta el grado de asociación de las variables con cada una de las tres *variables de autor* (*Pais_ categ*, *N_ antpolicial*, *Edad_ cat*). Básicamente se trata de ver qué variables de *modus operandi* tienen mayor asociación estadística para cada una de las tres *variables de autor*. Así, si dos variables tienen un alto grado de asociación estadística entre ellas, conocer el valor de una de ellas nos dará información sobre la otra y viceversa.

Una vez visto el concepto, falta concretar cómo calcularemos este grado de asociación entre variables. En nuestro caso se ha elegido el conocido estadístico V de Cramer. Este estadístico nos permite medir la asociación entre variables categóricas ya sean booleanas, ordinales, o combinaciones de ellas. Esto era un requisito imprescindible a la hora de elegir el método ya que todas las variables de nuestra base de datos son de este tipo. Además el estadístico proporciona un valor entre 0 y 1 independientemente de las dos variables que estemos comparando, siendo el valor 0 nula asociación estadística y 1 máxima asociación (presentan exactamente los mismos valores). Esto nos permite comparar las relaciones entre ellas y buscar las que presenten un mayor valor sin necesidad de normalizar.

La fórmula matemática de este estadístico es la siguiente:

$$V = \sqrt{\frac{\chi^2}{N(min[r,c]-1)}}$$

Donde χ^2 es el estadístico Chi-cuadrado, N el número total de elementos de la muestra a analizar, y r y c las filas y columnas de la tabla de contingencias usada para realizar la comparación de variables.

Para su cálculo en *R* se dispone de la función *cramersV()* a la que hay que pasar como argumento una tabla de contingencias de las dos variables a comparar previamente calculada. Una vez calculadas las V de Cramer para cada par *variable de autor-variable modus operandi* se han ordenado. Los resultados obtenidos se pueden ver en las figuras 4.2, 4.3 y 4.4, correspondiendo respectivamente a los valores para *Pais_ categ*, *N_ antpolicial*, *Edad_ cat*.

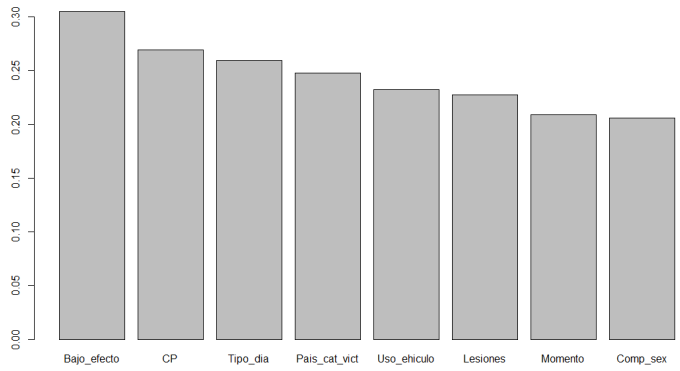

Figura 4.2: V de Cramer para las variables más asociadas con la variable de autor *Pais_ categ*

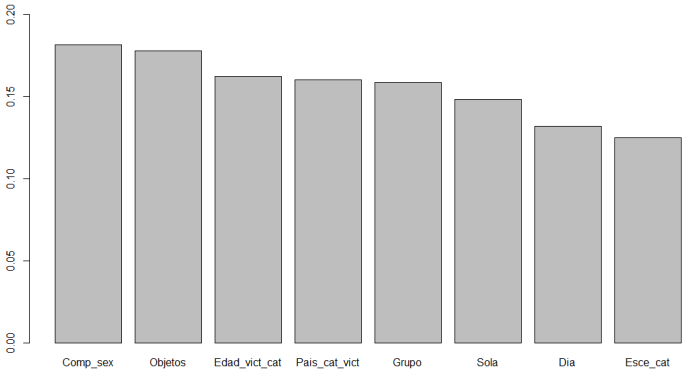

Figura 4.3: V de Cramer para las variables más asociadas con la variable de autor *N_ antpolicial*

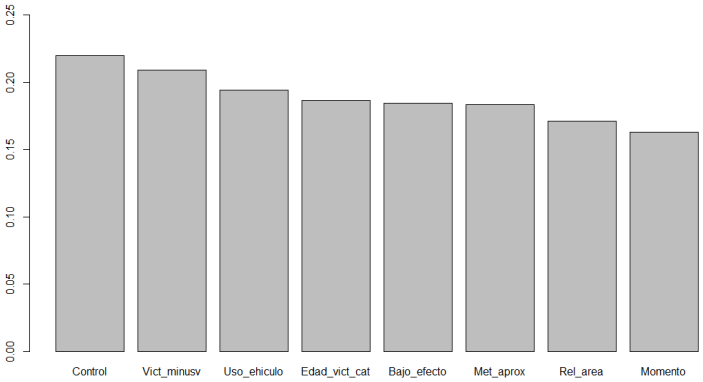

Figura 4.4: V de Cramer para las variables más asociadas con la variable de autor *Edad_ cat*

Estas asociaciones serán de especial utilidad a la hora de realizar entrenamientos con agrupación separado, ya que nos dan una idea de las variables de *modus operandi* que nos proporcionan mas información sobre la *variable de autor* que queremos predecir.

4.3. Sistemas de referencia

En todos los siguientes experimentos estaremos midiendo el rendimiento de las predicciones realizadas, a continuación se explican los sistemas de medida y referencia que se han usado para

ello.

De manera intuitiva, la primera forma de medida de rendimiento en la que se podría pensar es el porcentaje de acierto sobre el total de predicciones, también llamado precisión o *accuracy* en inglés. Sin embargo este sistema de medida puede resultar bastante engañoso en el problema al que nos enfrentamos, ya que a la hora de medir el rendimiento es conveniente tener en cuenta la variable que nos disponemos a predecir, sus posibles valores y sobre todo su distribución como veremos a continuación.

Por ejemplificar lo anteriormente descrito, imaginemos dos variables aleatorias que toman dos posibles valores (0 y 1) con distribución 80 %-20 % y 50 %-50 % respectivamente. Mientras que en el primer caso podremos obtener un 80 % de precisión con tan solo predecir siempre el valor 0, en el segundo esta precisión se reduce al 50 %. Si tomáramos como medida del rendimiento la precisión podríamos considerar que la primera se está prediciendo mejor que la segunda, sin embargo, desde nuestro punto de vista esto no es correcto ya que teniendo en cuenta el caso de la predicción trivial ambas se estarían prediciendo al mismo rendimiento. A este umbral que nos proporciona el valor de la predicción trivial (la más común) de cada variable lo denominamos línea base (o *baseline*).

Teniendo en cuenta lo anterior, consideraremos que una RB es válida prediciendo una variable si la precisión supera la línea base para esa variable. Para medir esto de tal forma que queden datos positivos en caso de superarla y negativos en caso de no hacerlo, usaremos como sistema de referencia la línea base y expresaremos los rendimientos de predicción (precisiones) en tanto por ciento relativo sobre ella.

4.4. Entrenamientos con agrupación simultánea

En esta sección explicaremos detalladamente y comentaremos los resultados obtenidos de los experimentos realizados con entrenamiendo con agrupación simultánea. Recordemos que en estos casos la RB a entrenar contiene las tres *variables de autor* a predecir.

En estos experimentos hemos utilizado la selección de variables por distribución, ya que las variables que consideramos útiles van a ser aquellas que ofrezcan bastante información, puesto que tenemos las tres variables de autor incluidas. Recordemos que para medir la distribución de las variables se ha usado la entropía normalizada.

4.4.1. Configuración y protocolo de entrenamiento

El protocolo utilizado para llevar a cabo el entrenamiento en este caso ha sido el siguiente:

- Seleción de variables por número de valores perdidos (máximo 15 %) y por distribución (valores de entropía normalizada inferiores a 0.5).
- Se han realizado combinaciones de hasta 12 variables y se ha calculado la mejor RB desde el punto de vista de la verosimilitud para cada combinación de variables.
- Se ha utilizado un límite máximo de complejidad de la red a entrenar. Este límite lo hemos definido mediante el uso de un factor de complejidad. La complejidad máxima de una red a entrenar será igual al número de variables de dicha red multiplicado por el factor de complejidad. Así, redes con más variables permitirán complejidades mayores para un mismo factor de complejidad.

- En cuanto al entrenamiento, por limitaciones computacionales se ha usado un orden preestablecido de los nodos. Colocando en primer lugar las *variables de autor* y después las variables de *modus operandi*, estas últimas ordenadas por su valor de entropía de mayor a menor (este orden puede verse en la gráfica 4.1). De esta forma conseguiremos que las variables a predecir queden en la zona inferior de la red (serán nodos hijos) y facilitaremos su inferencia.

- Para el aprendizaje/predicción se ha utilizado un algoritmo de validación cruzada con tamaño de *fold* 20 (escenario realista).

- En cuanto a la predicción, se ha realizado de siete maneras distintas, que se explicarán en cada uno de los siguientes apartados.

Como se ha explicado previamente, para comparar las distintas redes y medir el rendimiento de las predicciones, se ha utilizado la precisión (porcentaje de acierto) relativa sobre la línea base de cada variable.

Las gráficas que se mostrarán en los siguientes apartados reflejan la precisión de las distintas redes entrenadas. En el eje de las x se representará el ya explicado factor de complejidad y en el eje de las y la precisión relativa a la línea base (en porcentaje). Las distintas trazas corresponderán al tamaño de la red en cuestión. Como veremos, un aumento de la complejidad no tiene por qué conllevar una mejora en el rendimiento de la red, ya que muchas veces se llega a sobreajuste.

A continuación se exponen los resultados obtenidos para la predicción de cada una de las *variables de autor* en este contexto.

4.4.2. Predicción de variables de autor de una en una

En primer lugar se ha intentado predecir las *variables de autor* de una en una. Esto es, una vez entrenada la red en cada iteración de la validación cruzada, la red intentará predecir una sola de las *variables de autor*, teniendo disponibles todos los demás datos (siempre y cuando no haya un valor perdido).

País del agresor

Como podemos observar en la figura 4.5 se producen algunas mejoras puntuales sobre la línea base (representada como una línea negra en el 0). Esto puede ser debido a que algunas de las variables con mayor valor de V de Cramer para *Pais_ categ* están incluidas en la selección por distribución realizada para esta prueba.

Sorprende que las redes de solo cuatro y cinco nodos mejoren la línea base. Este hecho puede ser debido a que la variable *Pais_ categ_ vict*, con la que tiene un gran grado de asociatividad, es la primera en incluirse a la red con la selección por distribución llevada a cabo.

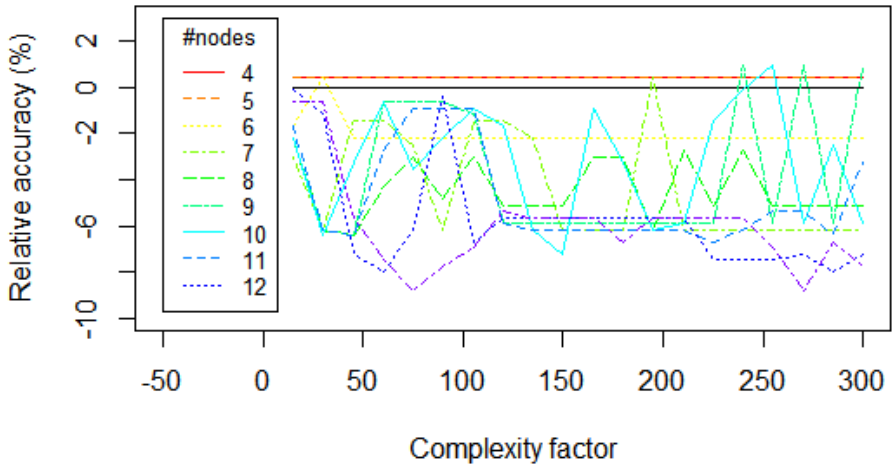

Figura 4.5: Precisión sobre línea base para la variable *Pais_ categ* con agrupamiento simultáneo, y tamaño de *fold* 20

Edad del agresor

En la figura 4.6 observamos que en este caso no se ha producido ninguna mejora en la predicción. De hecho vemos que los resultados son bastante malos. Es posible que la causa sea que no hay suficientes variables que aporten información para esta *variable de autor*, ya que ninguna de las variables con mayor valor de V de Cramer para *Edad_ cat* están incluidas en la selección por distribución realizada.

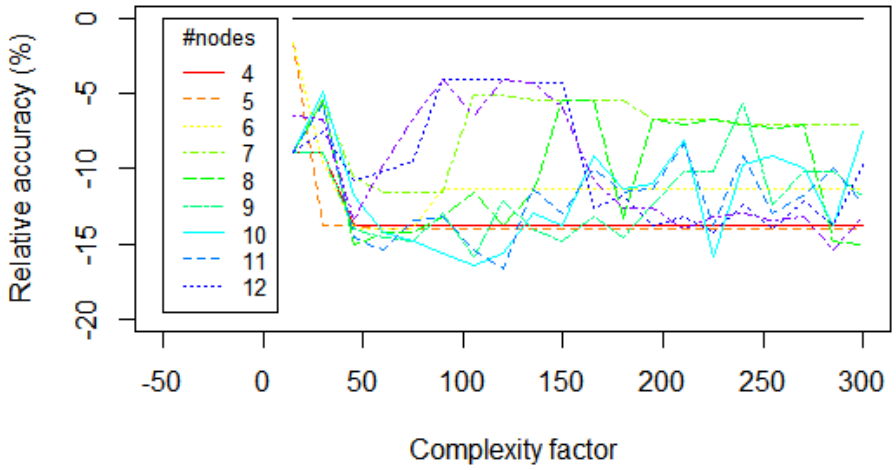

Figura 4.6: Precisión sobre línea base para la variable *Edad_ cat* con agrupamiento simultáneo, y tamaño de *fold* 20

Antecedentes policiales del agresor

En la figura 4.7 podemos ver los resultados obtenidos en el caso de predecir la variable *N_ antpolicial*. Los resultados son bastante malos, no solo no mejorando la línea base en ningún caso sino obteniendo resultados por debajo del 20 % de ésta. Esto puede ser debido a que en estas redes no han sido incluidas las variables que contaban con un mayor valor de V de Cramer para esta variable ya que se hizo una selección de variables por distribución.

Sorprende el hecho de que, en general, el rendimiento empeora con el aumento de la complejidad.

Aunque sorprendente, este resultado es en cierta manera esperado, pues como veremos a lo largo de los siguientes experimentos, *N_ antpolicial* es la *variable de autor* más difícil de predecir en nuestros casos. Algo que podíamos intuir tras observar los bajos valores de V de Cramer que se habían obtenido para esta variable en general.

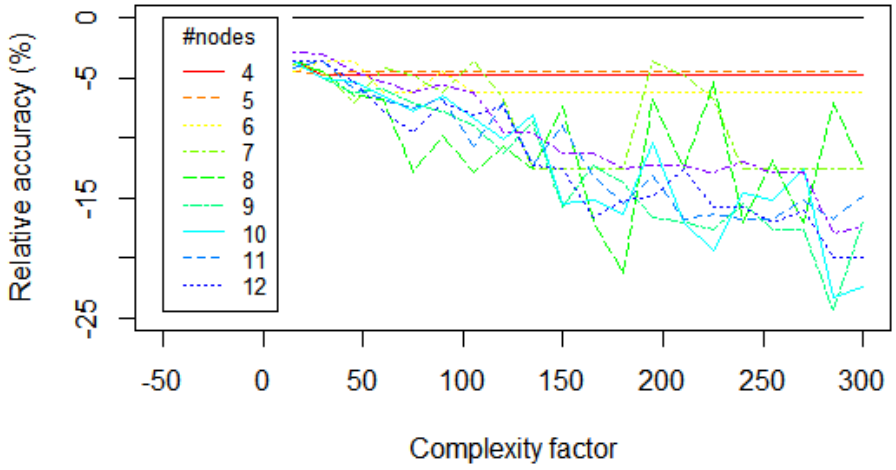

Figura 4.7: Precisión sobre línea base para la variable *N_ antpolicial* con agrupamiento simultáneo, y tamaño de *fold* 20

4.4.3. Predicción de variables de autor agrupadas de dos en dos

En segundo lugar se han intentado predecir las *variables de autor* agrupándolas de dos en dos. Esto es, una vez entrenada la red en cada iteración de la validación cruzada, la red intentará predecir dos de las *variables de autor*, teniendo disponible el valor de la tercera (siempre y cuando no sea un valor perdido).

País y edad del agresor simultáneamente

En las gráficas 4.8 y 4.9 podemos observar respectivamente los resultados para *Pais_ categ* y *Edad_ cat*. Comprobamos que ambas son muy similares a las obtenidas prediciéndolas por separado (correspondientes a las figuras 4.5 y 4.6 respectivamente) pero con resultados ligeramente peores. De hecho, para la predicción de *Pais_ categ* ya no mejora la línea base en ningún caso.

Cabe destacar que en ambos casos empeora de forma notable para las redes de pocos nodos. Es posible que esto sea debido a que al predecir las dos simultáneamente estamos perdiendo el valor de la otra a la hora de predecir. Esto es más notable cuando tenemos un número reducido de valores con los que predecir (pocos nodos).

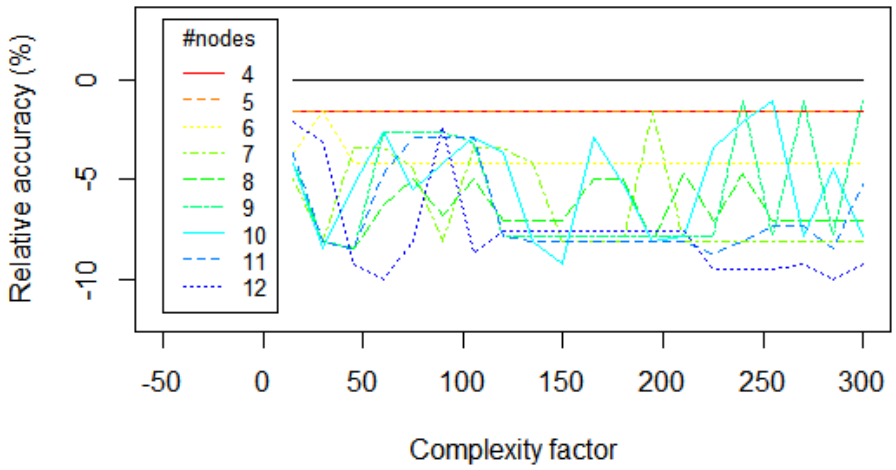

Figura 4.8: Precisión sobre línea base para la variable *Pais_ categ* con agrupamiento simultáneo, predicción simultánea con *Edad_ cat* y tamaño de *fold* 20

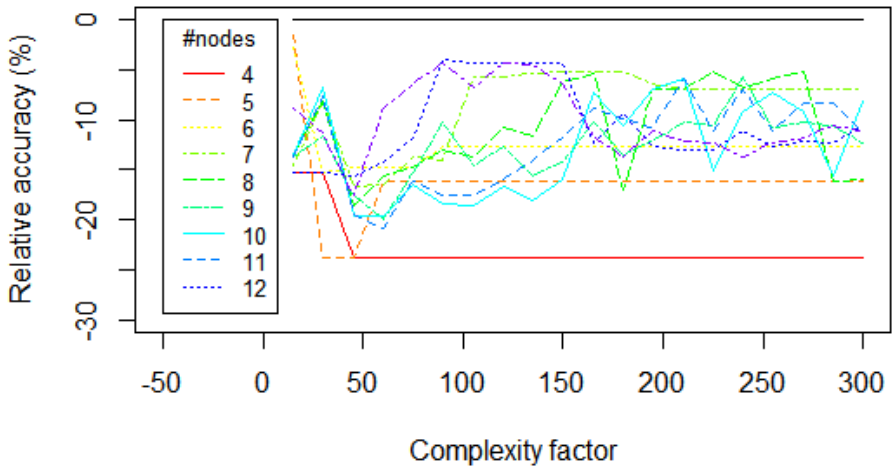

Figura 4.9: Precisión sobre línea base para la variable *Edad_ cat* con agrupamiento simultáneo, predicción simultánea con *Pais_ categ* y tamaño de *fold* 20

Edad y antecedentes del agresor simultáneamente

En las gráficas 4.10 y 4.11 podemos observar respectivamente los resultados para *N_ antpolicial* y *Edad_ cat*. Respecto a estas dos predicciones vemos que en ningún momento mejoran la línea base y no se observan cambios significativos entre ellas y las que se han obtenido previamente prediciendo por separado (figuras 4.7 y 4.6)

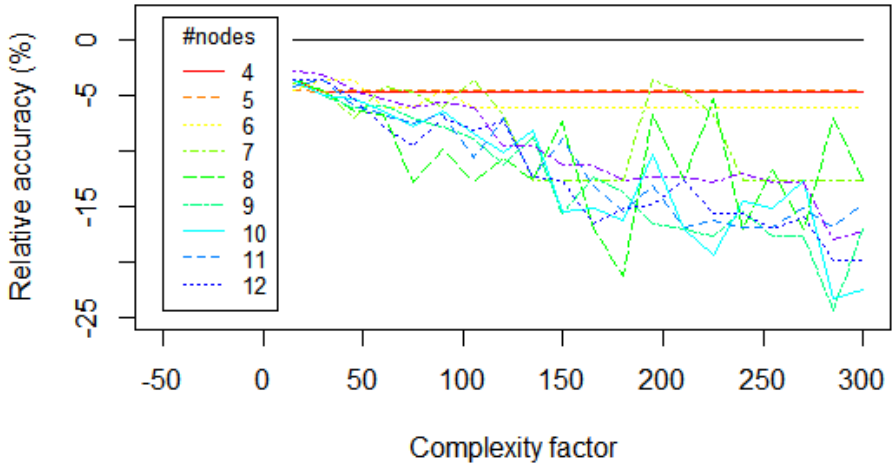

Figura 4.10: Precisión sobre línea base para la variable *N_ antpolicial* con agrupamiento simultáneo, predicción simultánea con *Edad_ cat* y tamaño de *fold* 20

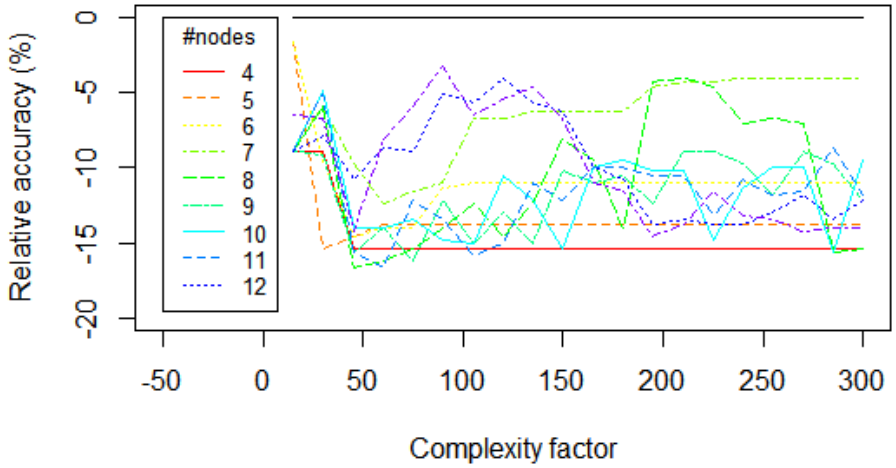

Figura 4.11: Precisión sobre línea base para la variable *Edad_ cat* con agrupamiento simultáneo, predicción simultánea con *N_ antpolicial* y tamaño de *fold* 20

País y antecedentes del agresor simultáneamente

En las gráficas 4.12 y 4.13 podemos observar respectivamente los resultados para *N_ antpolicial* y *Pais_ categ*. En ambos casos podemos ver ligeros cambios con respecto a las figuras 4.7 y 4.5 que predicen estas mismas variables de forma individual.

En el caso de *N_ antpolicial* se ha experimentado una mejora notable, pasando en algún punto de -25 % a -15 % de precisión relativa a la línea base. Pese a esta mejora los resultados siguen estando muy por debajo de la línea base.

En cuanto al caso de *Pais_ categ* las predicciones son ligeramente inferiores, no llegando en ningún caso a superar la línea base.

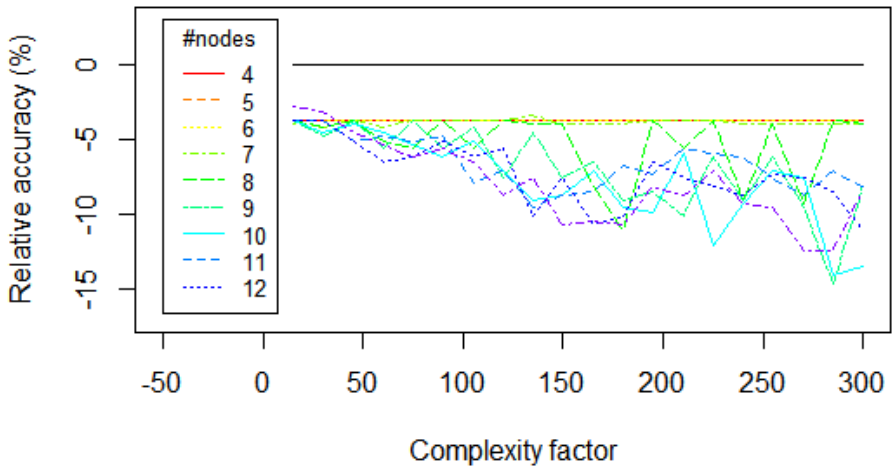

Figura 4.12: Precisión sobre línea base para la variable *N_ antpolicial* con agrupamiento simultáneo, predicción simultánea con *Pais_ categ* y tamaño de *fold* 20

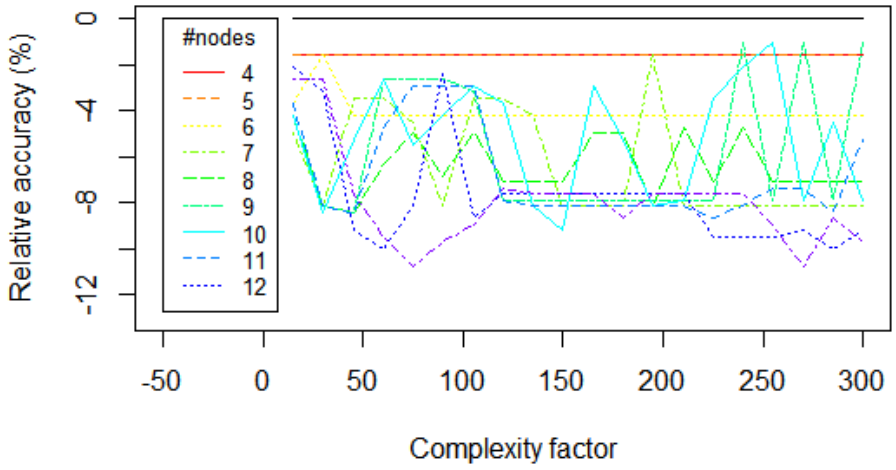

Figura 4.13: Precisión sobre línea base para la variable *Pais_ categ* con agrupamiento simultáneo, predicción simultánea con *N_ antpolicial* y tamaño de *fold* 20

4.4.4. Predicción de las tres variables de autor simultáneamente

En tercer y último lugar para esta configuración se han intentado predecir las tres *variables de autor* simultáneamente. Esto es, una vez entrenada la red en cada iteración de la validación cruzada, la red intentará predecir las tres variables de autor sin contar para la predicción con sus datos.

Siguiendo el razonamiento dado a los resultados previos, en este caso podemos esperar si cabe un empeoramiento aún mayor de las redes con un número reducido de nodos, ya que estamos denegando aún más información que en las pruebas anteriores.

En las gráficas 4.14, 4.15 y 4.16 podemos observar respectivamente los resultados para *N_ antpolicial* ,*Pais_ categ* y *Edad_ cat*. Comprobamos que se tratan de datos muy similares a los obtenidos prediciendo de dos en dos.

El cambio más notable lo vemos, como habíamos intuido, en las redes con menor número de nodos y más concretamente para la variable *Edad_ cat* (Figura 4.16), que tiene incluso peores resultados que en las predicciones de dos en dos.

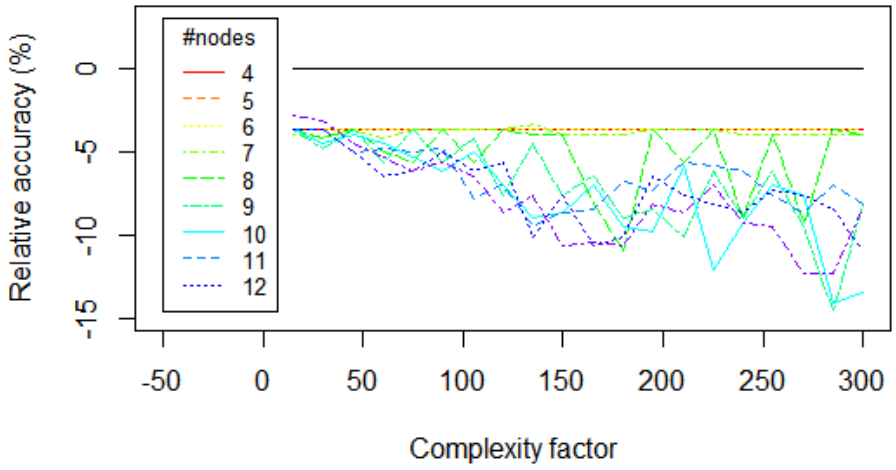

Figura 4.14: Precisión sobre línea base para la variable *N_ antpolicial* con agrupamiento simultáneo, predicción simultánea de las tres variables y tamaño de *fold* 20

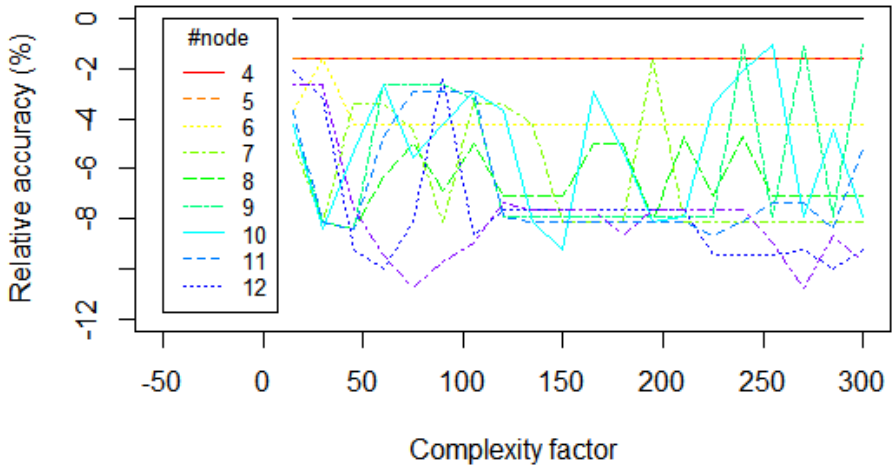

Figura 4.15: Precisión sobre línea base para la variable *Pais_ categ* con agrupamiento simultáneo, predicción simultánea de las tres variables y tamaño de *fold* 20

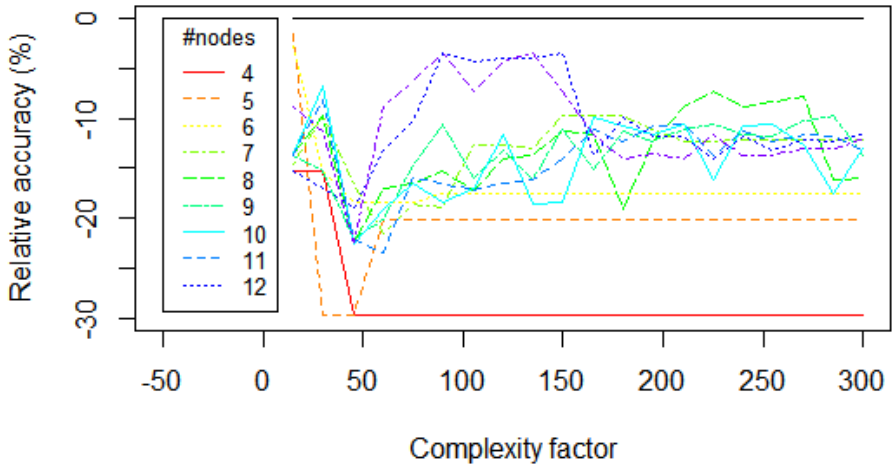

Figura 4.16: Precisión sobre línea base para la variable *Edad_cat* con agrupamiento simultáneo, predicción simultánea de las tres variables y tamaño de *fold* 20

4.4.5. Conclusión

Si bien hemos comprobado que los resultados con agrupamiento simultáneo y selección de variables por distribución no son buenos en general, se ha conseguido superar la línea base en uno de los casos probados.

Como hemos visto, el caso en el que se ha superado la línea base es la predicción de la *variable de autor Pais_categ* individualmente. Hemos comprobado que en la selección de variables por distribución utilizada en estas pruebas se incluyen variables más relacionadas con *Pais_categ* (Alto valor de V de Cramer) que con las otras dos *variables de autor*. Por ello se podría intuir que si establecemos un nuevo sistema de selección de variables teniendo en cuenta tanto asociatividad como capacidad informativa, podríamos seguir en una buena línea de trabajo y conseguir algún otro resultado positivo con agrupamiento simultáneo.

4.5. Entrenamientos con agrupación separada

En esta sección explicaremos detalladamente y comentaremos los resultados obtenidos de los experimentos realizados con entrenamiendo con agrupación separada. Recordemos que en estos casos entrenaremos una RB por cada una de las tres *variables de autor*.

En estos experimentos hemos utilizado la selección de variables por asociación. Para predecir de forma correcta una sola de las *variables de autor*, las variables de *modus operandi* que nos interesan (y las que más información nos van a proporcionar) son aquellas con las que tiene un mayor grado de asociación. Recordemos que el grado de asociación lo medíamos mediante el valor de V de Cramer de los pares *variable de autor-variable modus operandi* y que estos valores ya están calculados y se pueden consultar en las figuras 4.2, 4.3 y 4.4.

4.5.1. Configuración y protocolo de entrenamiento

El protocolo utilizado para llevar a cabo el entrenamiento en este caso ha sido el siguiente:

- Selección de variables por número de valores perdidos (máximo 15 %).
- Selección de variables por asociación para cada una de las *variables de autor*. Las variables seleccionadas para cada una pueden verse en las figuras 4.2, 4.3 y 4.4.
- Para cada conjunto de variables se han realizado combinaciones de hasta 10 variables y se han calculado las mejores RB desde el punto de vista de la verosimilitud.
- Se ha utilizado un límite máximo de complejidad de la red a entrenar. Este límite lo hemos definido mediante el uso de un factor de complejidad. La complejidad máxima de una red a entrenar será igual al número de variables de dicha red multiplicado por el factor de complejidad. Así, redes con más variables permitirán complejidades mayores para un mismo factor de complejidad.
- En cuanto al entrenamiento, por limitaciones computacionales se ha usado un orden preestablecido de los nodos. Siguiendo un razonamiento similar al utilizado en las pruebas de agrupamiento simultáneo se ha colocado primero la *variable de autor* correspondiente y después las variables de *modus operandi* ordenadas por su valor de V de Cramer de mayor a menor. De este modo, la *variable de autor* quedará en la zona inferior de la red facilitando su inferencia.
- Para el aprendizaje/predicción se ha utilizado un algoritmo de validación cruzada con tamaño de *fold* 20 (escenario realista).
- En cuanto a la predicción, se ha realizado por separado para cada *variable de autor*.

4.5.2. Resultados de las predicciones

En las figuras 4.17, 4.18 y 4.19 se pueden ver los resultados obtenidos para las predicciones de las *variables de autor* (*Pais_ categ*, *Edad_ cat* y *N_ antpolicial* respectivamente) para diferentes redes, modificando número de variables y factor de complejidad.

Lo primero que llama la atención es que para las tres variables se ha encontrado alguna red que mejora la línea base. Parece lógico pensar que si para cada variable seleccionamos las variables con las que tienen más relación se pueda obtener un resultado mejor que los obtenidos en las pruebas con agrupación simultánea, donde las variables de la red venían seleccionadas por un valor de distribución.

Se puede observar que la variable que tiene un mejor rendimiento predictivo en general es *Pais_ categ*, esto no es en realidad sorprendente, ya que habíamos visto que esta era la *variable de autor* que gozaba de mayor relación con las variables de *modus operandi* (valores más altos de las V de Cramer). Este hecho se puede observar en las figuras 4.2, 4.3 y 4.4.

Vemos en el caso opuesto, que la variable *N_ antpolicial* tiene los peores resultados en general. Su rendimiento está por encima de la línea base tan solo para redes de bajo factor de complejidad y número de nodos. Esto podría ser debido a la existencia de unas pocas variables con las que guarda una estrecha relación, pero sin embargo, en general, el valor de asociación con el resto de variables es bajo, hecho que se puede comprobar en la figura 4.2.

Como cómputo global para las tres variables podemos ver que un aumento de la complejidad o de la cantidad de nodos de la red no mejora la capacidad predictiva necesariamente. De hecho,

vemos que los mejores resultados los hemos obtenido en redes de 5-7 variables y factores de complejidad entre 20 y 50 en las tres variables.

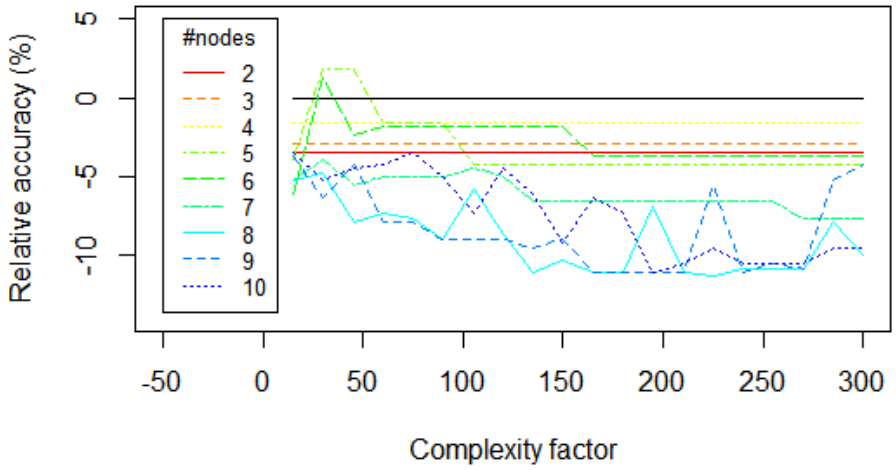

Figura 4.17: Precisión sobre línea base para la variable *Pais_ categ* con agrupamiento por separado y tamaño de *fold* 20

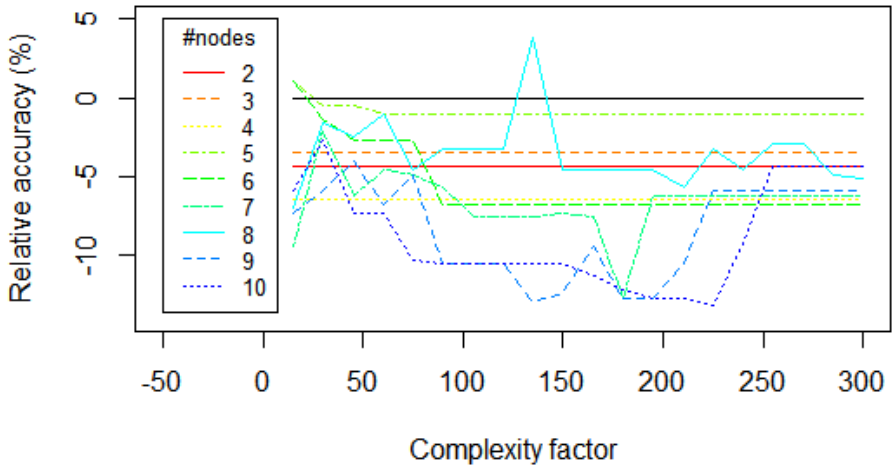

Figura 4.18: Precisión sobre línea base para la variable *Edad_ cat* con agrupamiento por separado y tamaño de *fold* 20

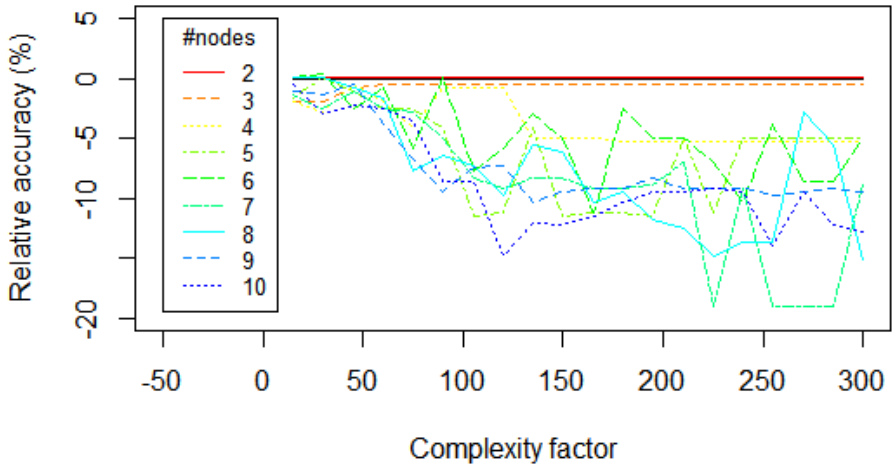

Figura 4.19: Precisión sobre línea base para la variable *N_ antpolicial* con agrupamiento por separado y tamaño de *fold* 20

A modo de ejemplo se ha realizado la predicción de *Pais_ categ* en un escenario optimista. Recordemos que en un escenario optimista se utiliza un tamaño de *fold* pequeño (en este caso 1) para favorecer el entrenamiento. Los resultados están reflejados en la figura 4.20 y como era de esperar son mucho mejores que los obtenidos en el escenario realista, superando la línea base en prácticamente todos los casos de 5, 6 y 7 nodos, en algunos de ellos por un amplio margen.

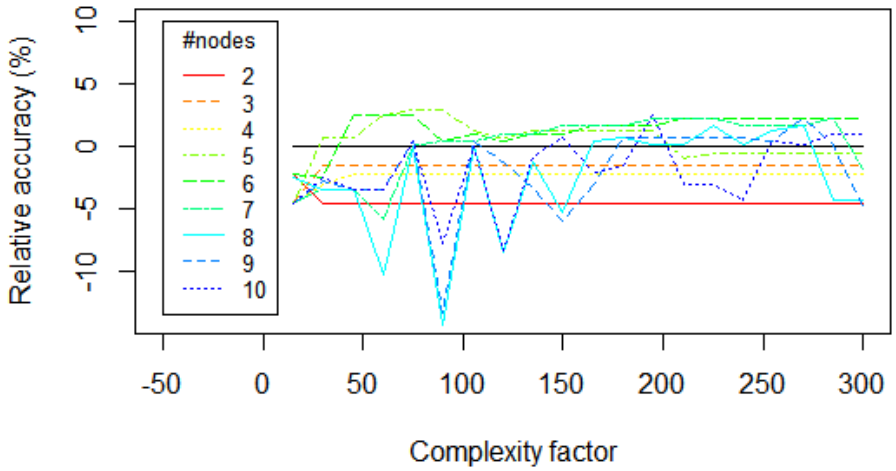

Figura 4.20: Precisión sobre línea base para la variable *Pais_ categ* con agrupamiento por separado y tamaño de *fold* 1

4.6. Predicción con agrupamiento de categorías poco frecuentes

Algunas de las variables con las que hemos trabajado en el presente proyecto presentan una distribución muy asimétrica de sus valores. En las figuras 4.21, 4.22 y 4.23 podemos ver algunos ejemplos de distribuciones de variables para las que existen algunos valores con muy pocas observaciones. Como se ha visto anteriormente, esto es un problema a la hora de obtener información válida de dichas variables (su valor de entropía es bastante alto).

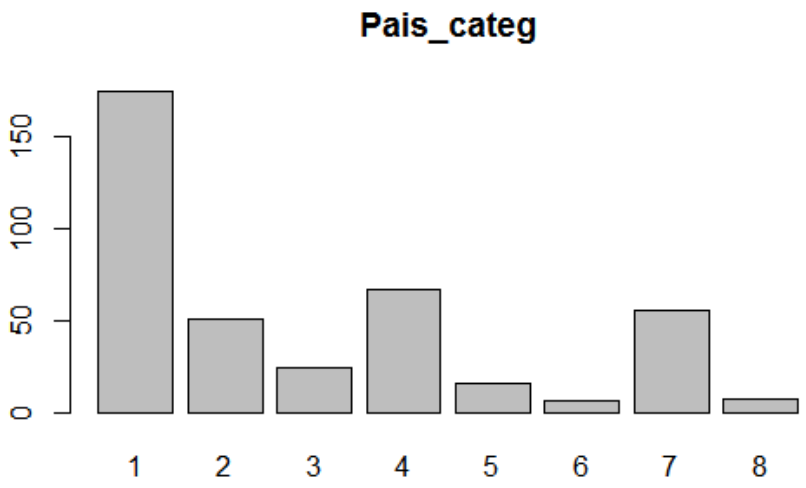

Figura 4.21: Distribución de la variable *Pais_ categ*

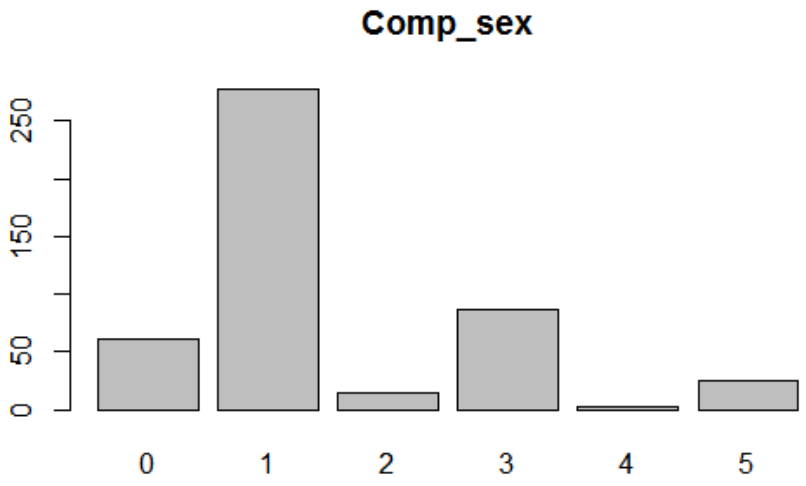

Figura 4.22: Distribución de la variable *Comp_ sex*

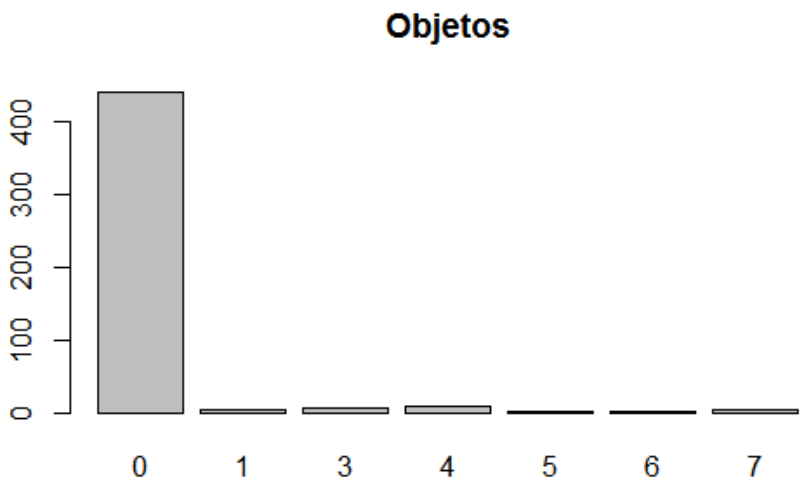

Figura 4.23: Distribución de la variable *Objetos*

Para solucionar este problema se ha pensado que podría ser buena idea agrupar algunos de los valores de estas variables, especialmente los que presentan menos observaciones, con el fin de conseguir una distribución más concentrada y disminuir así su entropía.

A continuación presentamos los resultados obtenidos tras agrupar algunos de estos valores en las mejores redes que hemos obtenido mediante el entrenamiento con agrupamiento por separado.

4.6.1. Red de País del agresor

La mejor red obtenida la hemos encontrado con 5 variables y factor de complejidad 30. En la figura 4.24 podemos ver la red con la que predeciremos tras los agrupamientos.

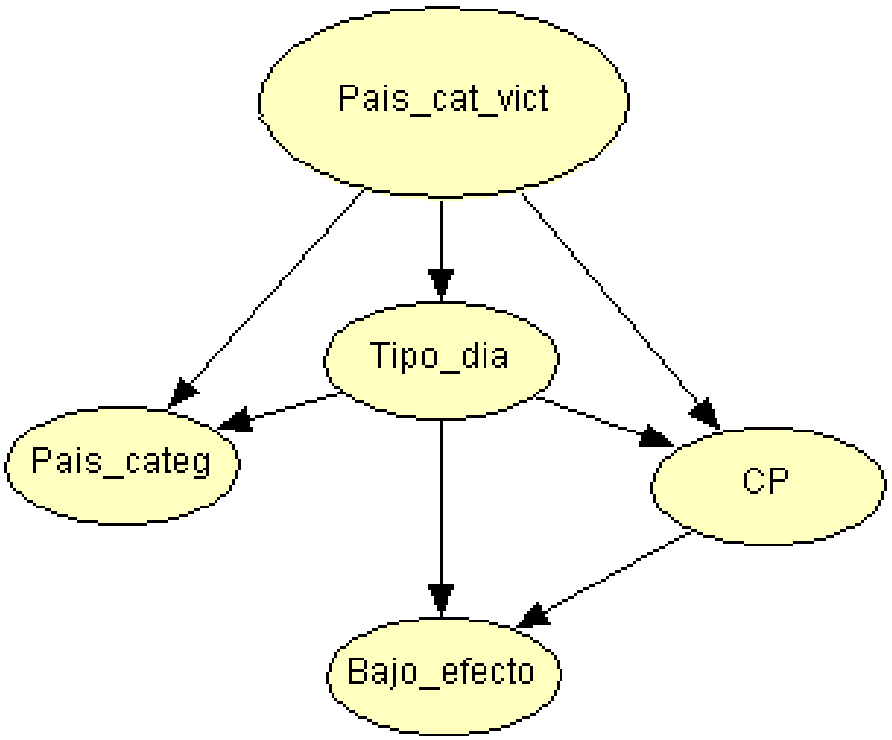

Figura 4.24: Mejor red obtenida para *Pais_ categ*

Tras estudiar las distribuciones de las 5 variables presentes en esta red se han realizado los siguientes agrupamientos de variables: valores 6-8 de *Pais_ categ*, valores 5-6-8 de *Pais_ categ*, valores 6-8 de *Pais_ vict*, valores 5-6-8 de *Pais_ categ*, valores 5-6-7-8 de *Pais_ categ*.

Para realizar las pruebas se ha probado a hacer los agrupamientos en distinto orden, obteniendo los mejores resultados para el orden que hemos presentado y para los que se obtienen los resultados de la gráfica de la figura 4.25.

La figura 4.25 muestra los resultados obtenidos. En el eje x se representan los agrupamientos realizados (en el orden que se ha indicado previamente) y en el eje y la precisión relativa a la precisión de la red sobre la que estamos trabajando. Como podemos observar solo se ha podido mejorar el rendimiento ligeramente cuando agrupamos los valores de la variable que estamos prediciendo *Pais_ categ*.

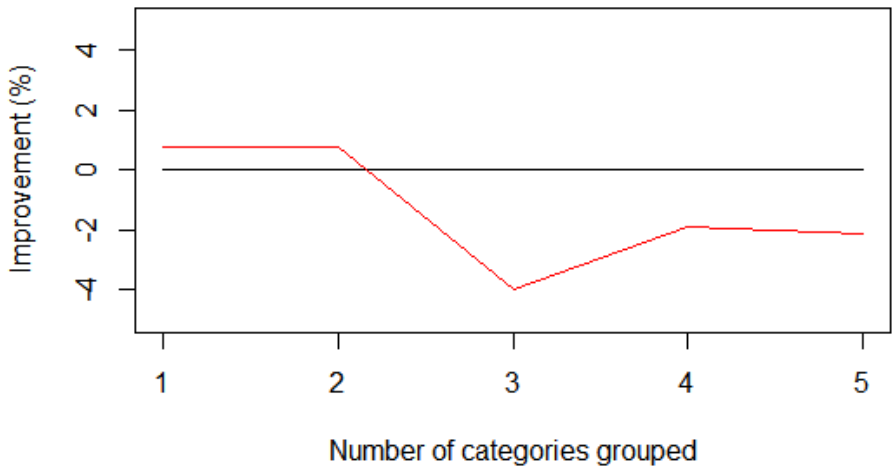

Figura 4.25: Rendimiento de la precisión de *Pais_ categ* en función del número de agrupamientos

4.6.2. Red de Edad del agresor

La mejor red obtenida la hemos encontrado con 8 variables y factor de complejidad 135. En la figura 4.26 podemos ver la red con la que predeciremos tras los agrupamientos.

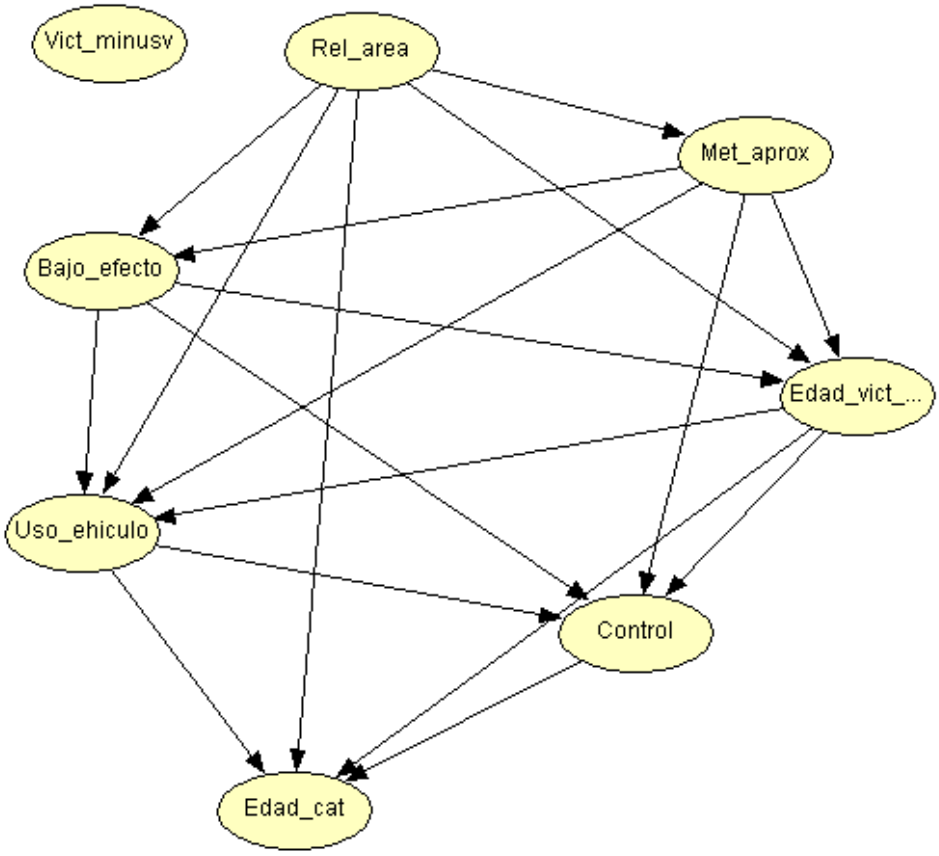

Figura 4.26: Mejor red obtenida para *Edad_ cat*

Tras estudiar las distribuciones de las 8 variables presentes en esta red se han realizado los siguientes agrupamientos de variables: valores 1-5 de *Edad_ cat*, valores 1-2 de *Control*, valores 1-5 de *Edad_ vict_ cat*, valores 1-2 de *Met_ aprox*, valores 1-4 de *Rel_ Area*.

Para realizar las pruebas se ha probado a hacer los agrupamientos en distinto orden, obteniendo los mejores resultados para el orden que hemos presentado y para los que se obtienen los resultados de la gráfica de la figura 4.27.

En este caso no hemos conseguido mejora de predicción mediante el agrupamiento de los valores poco frecuentes. Esto denota una debilidad estadística del modelo respecto a los valores con pocos datos.

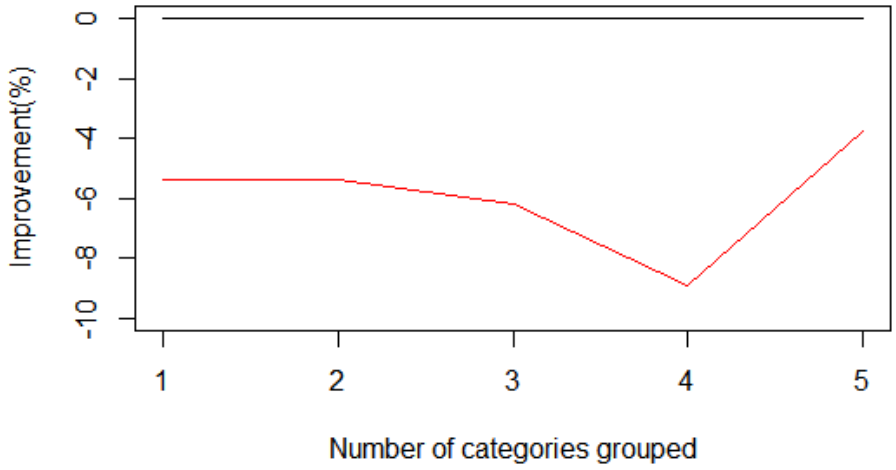

Figura 4.27: Rendimiento de la precisión de *Edad_ cat* en función del número de agrupamientos

4.6.3. Red de Antecedentes policiales del agresor

La mejor red obtenida la hemos encontrado con 6 variables y factor de complejidad 90. En la figura 4.28 podemos ver la red con la que predeciremos tras los agrupamientos.

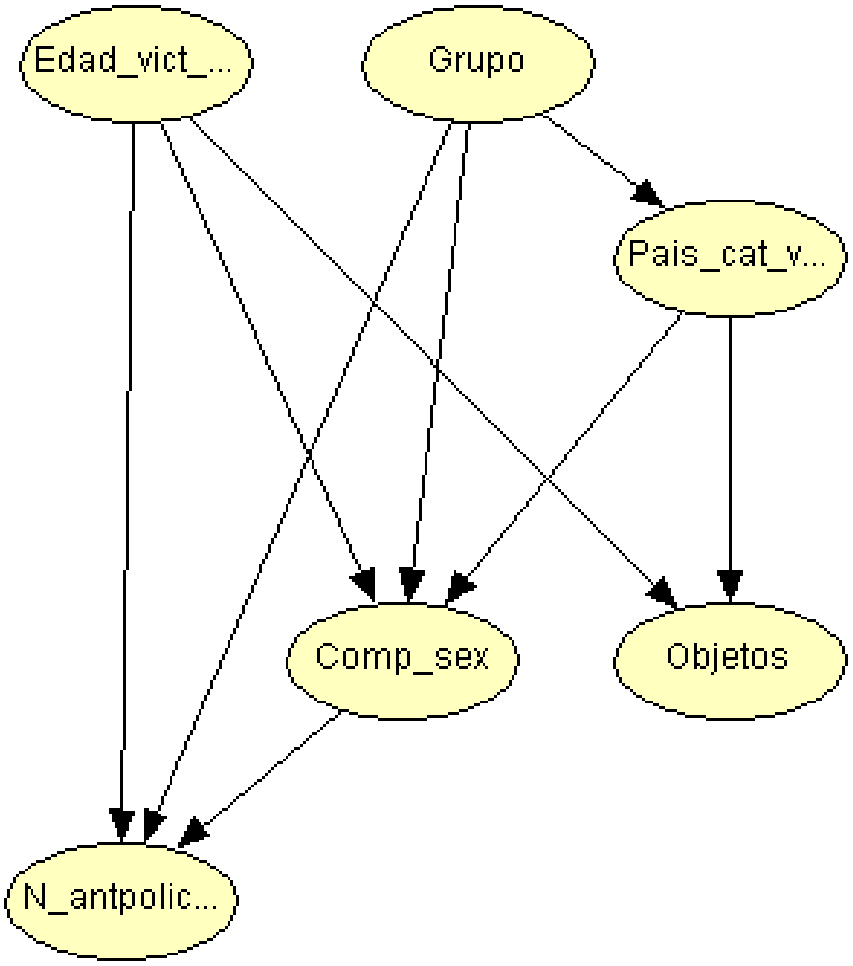

Figura 4.28: Mejor red obtenida para *N_ antpolicial*

Tras estudiar las distribuciones de las 6 variables presentes en esta red se han realizado los siguientes agrupamientos de variables: valores 1-2 de *N_ antpolicial*, valores 2-4 de *Comp_ sex*, valores 5-6-1-7 de *Objetos*, valores 1-5 de *Edad_ Vict_ Cat*, valores 5-6-8 de *Pais_ cat_ vict*.

Para realizar las pruebas se ha probado a hacer los agrupamientos en distinto orden, obteniendo los mejores resultados para el orden que hemos presentado y para los que se obtienen los resultados de la gráfica de la figura 4.29.

Nos encontramos ante unos resultados similares a los obtenidos en la figura 4.25 donde el agrupamiento de los valores poco frecuentes de la variable a predecir mejora la capacidad predictiva de la red.

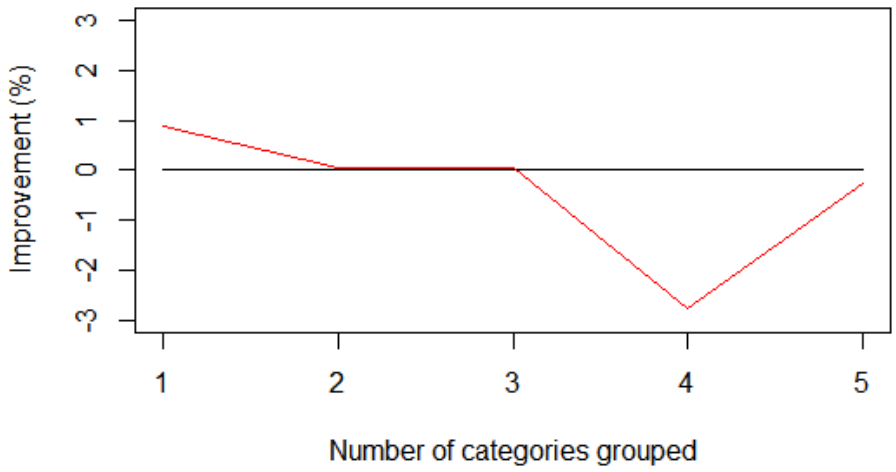

Figura 4.29: Rendimiento de la precisión de *N_ antpolicial* en función del número de agrupamientos

4.6.4. Conclusión

Tras observar los resultados obtenidos en las agrupaciones cabe preguntarse si es mejor acertar un mayor número de predicciones o tener predicciones más específicas, aunque el número correcto de éstas sea inferior.

Tras las pruebas realizadas en este apartado hemos comprobado que somos capaces de mejorar la capacidad predictiva de las redes previas. La cuestión a la que nos enfrentamos ahora es el precio a pagar por ello, ya que imponer un agrupamiento en sus valores hace que la variable nos proporcione información mucho menos específica.

En cualquier caso, estos resultados demuestran que el modelo sobre el que trabajamos y con los valores que disponen de pocas observaciones tiene una robustez limitada, ya que no hemos conseguido mejorar en ningún caso agrupando valores de variables de *modus operandi*.

5

Conclusiones y trabajo futuro

El grado de cumplimiento con los objetivos iniciales del proyecto ha sido en general bastante satisfactorio. A continuación se expone en más detalle los principales logros y conclusiones a las que se ha llegado en este trabajo.

1. Hacer un estudio teórico de los modelos gráficos probabilísticos, más en concreto de las RB y su aplicación en problemas parecidos a los que nos enfrentaremos en el trabajo.

   - Gracias al estudio teórico sobre Redes Bayesianas realizado ha sido posible tener una idea previa aproximada sobre qué esperar de cada prueba realizada. Esto ha permitido, por un lado, detectar errores a la hora de la implementación de las pruebas cuando se han obtenido resultados alejados de lo que cabía esperar, y por otro lado, la selección de la configuración óptima para los escenarios de prueba.

2. Generación y entrenamiento de Redes Bayesianas en R utilizando la base de datos generada por el IFCS-UAM.

   - Gracias al lenguaje de programación *R* y la librería catNet se han implementado métodos de entrenamiento de Redes Bayesianas siguiendo entre otras las siguientes estrategias:
     I) Distintas formas de selección de variables.
     II) Distintos esquemas de agrupación de variables.
     III) Agrupación de valores poco frecuentes.
     IV) Coste computacional
   - En este sentido, se ha demostrado la posibilidad de crear un modelo probabilístico basado en Redes Bayesianas que refleja de forma realista y correcta los datos contenidos en la base de datos de partida. Gracias a ello hemos sido capaces de predecir por encima de la línea base en algunos de los casos probados.

3. Estudio empírico de las capacidades predictivas de las RB implementadas utilizando los datos del IFCS-UAM.

   - Se han llevado a cabo pruebas sobre las capacidades predictivas de las distintas Redes Bayesianas entrenadas. Para ello, se han realizado numerosas gráficas que reflejan el rendimiento predictivo en tanto por ciento sobre la línea base para distintos valores de complejidad y número de variables en cada configuración de entrenamiento y predicción.

- Para cada uno de los escenarios de pruebas de los que partíamos se ha conseguido mejorar la línea base en alguna de sus configuraciones. Puesto que este era el objetivo propuesto inicialmente, esta mejora denota un buen análisis teórico previo a la hora de establecer o seleccionar los distintos escenarios de prueba y sus configuraciones.
- Los mejores resultados se han obtenido para redes con agrupamiento por separado y agrupamiento de los valores poco frecuentes de la *variable de autor* a predecir en cada caso. Si bien es cierto que las capacidades predictivas de las redes en estos casos son superiores, los resultados que se obtienen de ellos son sin embargo menos específicos. Esto abre una discusión sobre qué es preferible que los expertos analicen.

4. Estudio de las RB generadas y su funcionalidad utilizando la herramienta HuginLite.

   - Para poder llevar a cabo el estudio de las Redes Bayesianas con HuginLite se ha implementado una función en R que adapta los datos al formato correspondiente.
   - El análisis de redes mediante HuginLite ha sido realmente importante en algunos casos en los que no se obtenían los resultados esperados. Esto ha permitido estudiar el comportamiento de la red en cuestión y detectar el error que se estaba cometiendo.

Como posible trabajo futuro sería interesante seguir realizando pruebas con algún tipo de esquema de selección de variables híbrido (por asociación y distribución) y con agrupación simultánea, ya que hemos visto que en algunos casos se ha superado la línea base siguiendo esta línea de trabajo. Los entrenamientos de agrupación simultánea de las *variables de autor* tienen especial importancia ya que son los que reflejan de forma más realista el modelo sobre el que queremos predecir al contener todas las variables. Asimismo podría ser un buen motivo de trabajo futuro realizar pruebas con algo más de complejidad, aunque para esto será necesario aumentar la capacidad computacional de la que se dispone.

Glosario de acrónimos

- **FCSE**: Fuerzas y Cuerpos de Seguridad del Estado
- **ICFS-UAM**: Instituto de las Ciencias Forenses y la Seguridad de la Universidad Autónoma de Madrid
- **MG**: Modelo Gráfico
- **RB**: Red bayesiana
- **BIC**: Criterio de información bayesiano *(Bayesian Information Criterion)*
- **AIC**: Criterio de información de Akaike *(Akaike Information Criterion)*
- **LOO**: Dejar uno fuera *(Leave One Out)*
- **EM**: Algoritmo esperanza-maximización *(Expectation Maximization)*
- **MLE**: Estimación de máxima verosimilitud *(Maximum-Likelihood estimation)*

Bibliografía

[1] Nir Friedman Daphne Koehler. "*Graphical Models in a Nutshell*". *MIT Press*, 2009.

[2] Kevin P. Murphy. "*An introduction to Graphical Models*". 2001.

[3] Peter Salzman Nikolay Balov. "*Categorical Bayesian Network Inference*". 2017.

[4] D. G. Stork R. O. Duda, P. E. Hart. "*Pattern classification*". *John Wiley and Sons*, 2000.